DOCUMENT RESUME

ED 324 976 FL 018 990

AUTHOR	Stansfield, Charles W.; And Others
TITLE	Spanish-English Verbatim Translation Exam. Final Report.
INSTITUTION	Center for Applied Linguistics, Washington, D.C.
SPONS AGENCY	Federal Bureau of Investigation, Quantico, VA.
PUB DATE	30 Nov 90
NOTE	220p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	Content Validity; *English; Language Proficiency; *Language Tests; *Spanish; Test Construction; Test Items; Test Validity; *Translation
IDENTIFIERS	*Federal Bureau of Investigation; *Spanish English Verbatim Translation Exam

ABSTRACT

The development and validation of the Spanish-English Verbatim Translation Exam (SEVTE) is described. The test is for use by the Federal Bureau of Investigation (FBI) in the selection of applicants for the positions of Language Specialist or Contract Linguist. The report is divided into eight sections. Section 1 describes the need for the test, reviews the literature on the testing of translation ability, and discusses the development of translation skill level descriptions. Section 2 describes the multiple-choice and production sections of the SEVTE, scoring procedures and time limits. Sections 3 and 4 describe the development, trialing, and pilot testing. Section 5 describes the design and validation study, which included members of the FBI, Houston Police Department, and professional translators. Section 6 presents statistics on the scores of the subjects, and analyzes the reliability of each SEVTE section. Section 7 discusses content validity. Section 8 describes the equating of the two parallel forms, and the establishment of a cut score on the SEVTE multiple-choice section. Appended materials include sample test items, administration instructions, scoring guidelines, the FBI/Center for Applied Linguistics Translation Skill Level Descriptions, questionnaires, and other data collection instruments. (Author/VWL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED324976

SPANISH - ENGLISH VERBATIM TRANSLATION EXAM

Final Report

by

Charles W. Stansfield

Mary Lee Scott

Dorry Mann Kenyon

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

Center for Applied Linguistics
1118 22nd St. N.W.
Washington, DC 20037

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G.R. Tucker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

November 30, 1990

FL018990

Abstract

This document describes the development and validation of the Spanish - English Verbatim Translation Exam (SEVTE) for use by the Federal Bureau of Investigation (FBI) in the selection of applicants for the positions of Language Specialist or Contract Linguist. The report is divided into eight sections. Section 1 describes the need for the test, reviews the literature on the testing of translation ability, and discusses the development of translation skill level descriptions. Section 2 describes the multiple-choice and production sections of the SEVTE, scoring procedures and time limits. Section 3 and 4 describe its development, trialing and pilot testing on translation students at Georgetown University. Section 5 describes the design of the validation study, which included 44 employees of the FBI, members of the Houston Police Department, and professional translators. Section 6 presents descriptive statistics on the scores of the above subjects, and analyses the reliability of each SEVTE section using traditional methods and Generalizeability theory. The results indicate that the SEVTE is quite reliable for a test that involves free response items. Section 7, the longest of the report, begins with a discussion of content validity. Subsequent subsections discuss the evidence for construct, criterion-related, convergent and discriminant validity based on the results of the validation study. The results indicate that the two SEVTE constructs, Accuracy and Expression, are interrelated, but measure different dimensions of translation ability. Section 8 describes the equating of the two parallel forms, and the establishment of a cut score on the SEVTE multiple-choice section, which can be used as a screening test. The 18 appendices include sample test items, administration instructions, scoring guidelines, the FBI\CAL Translation Skill Level Descriptions, questionnaires and other data-collection instruments.

Acknowledgements

A project of this magnitude could not have been carried out without the cooperation and assistance of many people. We are indebted to the following people for their help over the past two years, during which time this project was being carried out.

Marijke Walker, the contractor's technical representative at the FBI, arranged for meetings between CAL staff and FBI staff, arranged for the SEVTE to be administrered at FBI offices around the country, and worked as a colleague with us on the Translation Skill Level Descriptions. She provided important feedback at critical decision points during the project.

Ana Maria Velasco assisted in the development of CAL's proposal to the FBI, drafted the needs analysis questionnaire, wrote items for the SEVTE, assisted in the development of the scoring guidelines and the Translation Skill Level Descriptions, and scored the pretest versions of the SEVTE.

Stephanie Kasuboski performed ably as the CAL project coordinator over a six-month period while pretest versions were being developed. She drafted the examinee questionnaire that was used in the SEVTE trialing and analyzed the completed questionnaires.

Kathleen Marcos assisted in the writing of the CAL proposal and reviewed items for the pretest and final version of the SEVTE. She also provided clerical assistance, analyzed the returned questionnaires from the survey of translation needs, and supervised the pretest administration at Georgetown University.

Matilde Farren assisted in the development of the test items, and scored half of the validation study exams.

Agnes B. Werner assisted in the development and reviewing of test items and in the scoring of the pretests.

Carol Sparhawk assisted in the preparation of materials for the training of FBI raters and organized the appendices for this final report.

Laurel Winston and Elizabeth Franz provided clerical assistance on many occasions.

Katrine Gardner of the CIA Language School arranged for CIA Spanish language students to take the Multiple Choice section of the exam.

Olga Navarrete functioned as liaison between CAL and the FBI. She and other FBI staff members took and commented on pretest and final versions of the test.

In addition to the above, we would like to acknowledge the cooperation of the staff at FBI field offices in Albuquerque, El Paso, San Juan, Miami, Los Angeles, San Antonio, and San Diego. At each of these offices, administrators arranged for Special Agents, Language Specialists, Contract Linguists, and support personnel to have released time to take the exam and returned all test booklets to CAL.

We would also like to acknowledge the cooperation of the members of Houston Police Department who also took the exam.

At CAL, John Karl, Joy Peyton, and Peggy Seifert Bosco took and commented on pretest versions of the test.

Professor Lyle F. Bachman of the University of California at Los Angeles made helpful comments on an earlier version of the report.

We are grateful to the above individuals and to the many others who played a role in this project. Most especially, we are grateful to the FBI for awarding us a contract to develop this test and to carry out the research associated with its validation.

Abstract

This report describes the development and validation of the Spanish - English Verbatim Translation Exam (SEVTE). The SEVTE was developed by staff at the Center for Applied Linguistics (CAL) under contract with the Federal Bureau of Investigation (FBI). The SEVTE is designed to be a job relevant test of the ability to render a translation in English of a text written in Spanish. The report is divided into five sections, plus appendices.

Section 1 provides an introduction to the project and establishes a framework for the project. This section describes the groups that would potentially be given the test, the survey of the types of documents the FBI needs to have translated, the development of ILR skill level descriptions for translation, the nature of translation, and the emergence of the two constructs of translation ability that are measured by the SEVTE.

Section 2 provides a description of the test, which is divided into multiple choice and free response sections. The scoring of the test is also described and the computation of the total scores on two criteria, Accuracy and Expression, are discussed.

Sections 3 and 4 describe the development and pilot testing of the SEVTE and the successive revisions it underwent.

Section 5 describes the validation study that was conducted on the final version of the test. It discusses the test administration procedures, the sample, and the scoring of the

tests. For this study, 66 examinees took both forms of the SEVTE. The subjects were FBI Language Specialists, Special Agents, and support staff, as well as members of the Houston TX Police Department and employees of the Central Intelligence Agency.

Section 6 presents descriptive statistics on test performance from the validation study as well as a detailed analysis of the reliability of the test. Reliability analyses include internal consistency, product moment correlations, and generalizability coefficients.

Section 7 presents the discussion of the validity of the exam. For this study, additional data was collected from employee files in the form of independent measures of proficiency in Spanish and English, and scores on an earlier generation of FBI translation tests. Subjects also completed a self-rating of the ability to translate various types of FBI documents. A number of statistical analyses were performed on the data. The results establish the validity of the constructs measured and support the validity of the SEVTE for the screening, selection, and placement of FBI applicants and staff in positions requiring Spanish - English translation ability.

Section 8 of the report describes the development of a score conversion table, which can be used to convert scores on the SEVTE to an overall rating of translation proficiency on a 0 to 5 scale.

Eighteen appendices follow the body of the report. These provide additional data and information relating to matters discussed in the text.

Table of Contents

Acknowledgements 2

Abstract 5

Table of Contents 8

List of Appendices 10

List of Tables 11

1. Introduction 12
  1.1. Need for the Test 12
  1.2. Intended Use 14
  1.3. FBI Translation Needs Survey 15
  1.4. FBI\CAL Translation Skill Level Descriptions . . . 16
    1.4.1. History 16
    1.4.2. Explanation of the Skill Level Descriptions 23
  1.5. The Nature of Translation Ability 27
    1.5.1. The Need to Define the Construct 28
    1.5.2. The Literature on Translation 29
    1.5.3. The Emergence of the Constructs 31

2. General Description 37
  2.1. Multiple Choice Section 37
    2.1.1. Format 37
    2.1.2. Test Taking 38
    2.1.3. Scoring Procedures 39
  2.2. Production Section 39
    2.2.1. Format 39
    2.2.2. Test Taking 40
    2.2.3. Scoring 40
      2.2.3.1. Words or Phrases in Sentences Items 40
      2.2.3.2. Sentence Translation Items 40
      2.2.3.3. Paragraph Translation Items 41
  2.3. Computation of Total Scores 42
  2.4. Use of Multiple Choice Section for Screening . . . 43

3. Development of the SEVTE 44
  3.1. Exam Forms 44
  3.2. Pilot Test Scoring Procedures 45

4. Trialing and Pilot Testing 48
  4.1. Trialing 48
  4.2. Pilot Testing 48
    4.2.1. Data Collection 49
    4.2.2. Results 50

4.2.3. Revisions 51

5. Validation Study 54
5.1. Overview 55
5.1.1. Test Administration Instructions 56
5.1.2. Questionnaires 56
5.1.3. Subjects 57
5.2. Scoring 59

6. Reliability 62
6.1. Multiple Choice Section: Descriptive Statistics and Reliability 62
6.2. Production Section: Descriptive Statistics and Reliability of the Accuracy Score 64
6.3. Production Section: Descriptive Statistics and Reliability of the Expression Score 69

7. Examining the Validity of the SEVTE 78
7.1. Content Validity 79
7.2. Construct Validity 87
7.3. Criterion-related Validity 90
7.3. Convergent/Discriminant Construct Validity 95
7.4.1. Convergent Validity 98
7.4.2. Discriminant Validity 105
7.4. Conclusions 109

8. Construction of Translation Skill Level Score Conversion Tables for the SEVTE 113
8.1. Overview 113
8.2. Determining Contributors to Expression and Accuracy Total Scores 114
8.3. Development of Raw Score to Scaled Score Conversion Tables 116
8.4. Using the Multiple Choice Section as a "Screen" . . 117

References 120

List of Appendices

Appendix A. Administration Instructions for SEVTE
Appendix B. Multiple Choice Section Instructions and Title Page
Appendix C. Production Section Test Instructions
Appendix D. Content Analysis of SEVTE MC Sections
Appendix E. Sentence Accuracy Scoring Guidelines
Appendix F. Paragraph Scoring Guidelines
Appendix G. Pilot Version of Sentence Scoring Grid
Appendix H. Pilot Version of Paragraph Scoring Grid
Appendix I. FBI/CAL Translation Skill Level Descriptions and Questionnaire
Appendix J. Background Proficiency Questionnaire, Given before Trialing
Appendix K. Exam Feedback Questionnaire, Multiple Choice and Production Sections (Trialing Version)
Appendix L. SEVTE Exam Feedback Questionnaire (Validation Study)
Appendix M. Pilot Questionnaire and Results on Language Background and Proficiency
Appendix N. Self-Assessment Questionnaire and Summary Report on Self-Assessment
Appendix O. Conversion Tables: Raw Score to TSL Score Expression and Accuracy
Appendix P. Memorandum on Total Score Conversion to FBI/CAL Equivalency Rating
Appendix Q. Survey of FBI Translation Needs
Appendix R. RFP Statement of Work

List of Tables

Table 1 SEVTE Multiple Choice Sections Total Pilot Sample . 50

Table 2 Descriptive Statistics for SEVTE MC1 and MC2 . . . 62

Table 3 KR-20 Reliability for MC1 and MC2 63

Table 4 Descriptive Statistics for SEVTE Accuracy 65

Table 5 Interrater Reliability of SEVTE Production Subsections and Production Total for Accuracy . . . 66

Table 6 Coefficient of Equivalence for SEVTE Accuracy Scores 67

Table 7 Variance Contributions of Raters and Forms to the SEVTE-Accuracy Total Score 68

Table 8 Estimated Generalizability Coefficients for the SEVTE-Accuracy Score using Different Groupings of Forms and Raters 69

Table 9 Descriptive Statistics for SEVTE Expression: Paragraphs Subsection 70

Table 10 Interrater Reliability of SEVTE Production Subscores and Production Total 71

Table 11 Coefficient of Equivalence for SEVTE Expression Scores 72

Table 12 Variance Contributions of Raters and Forms to the SEVTE-Expression Production Total Score 73

Table 13 Estimated Generalizability Coefficients for the SEVTE-Expression Production Score using Different Groupings of Forms and Raters 74

Table 14 Coefficient of Equivalence for SEVTE Expression Composite Scores 76

Table 15 Correlations between Mean Total Expression and Accuracy Scores 89

Table 16 Correlations of the SEVTE Scores with Overall Rating of Translation Ability 92

Table 17 Correlations of the SEVTE Scores with Other Available Measures 99

1. Introduction

This section of the report on the Spanish into English Verbatim Translation Exam (SEVTE) is intended to provide the reader with some appropriate background as a preliminary to a discussion of the test.

1.1. Need for the Test

The Federal Bureau of Investigation (FBI) is the Federal Government's principal agency responsible for investigating violations of federal statutes. The overall objective of the FBI is to investigate criminal activity and civil matters in which the Federal Government has an interest, and to provide the Executive Branch with information relating to national security. FBI activities include investigations into organized crime, white-collar crime, public corruption, financial crime, fraud against the Government, bribery, copyright matters, civil rights violations, bank robbery, extortion kidnaping, air piracy, terrorism, foreign counterintelligence, interstate criminal activity, fugitive and drug trafficking matters, and other violations of more than 260 federal statutes.

In all of the above areas of jurisdictional responsibility, it is likely that the FBI could be called upon to investigate a large number of cases that involve languages other than English. Because of this, it is understandable that the FBI is increasingly called upon to provide Special Agents and other employees who are proficient in a foreign language. All modes of communicative skills may be required. That is, FBI staff may need to be able to speak, understand, read or write the foreign language. They may also be required to provide oral interpretation or written translation. Often, they are called upon to provide a written summary in English of a foreign language conversation.

The need to assess employees' or potential employees' language skills can be satisfied in a number of ways. To measure the speaking skill, the FBI has used the Interagnecy Language Roundtable (ILR) Oral Proficiency Interview for many years. To measure the listening and reading skills, the FBI uses the Listening and Reading sections of the Defense Language Proficiency Test (typically version II), (Walker, et al., 1988). These exams are taken by applicants for the position of Special Agent Linguist,[1] Language Specialist, and Contract Linguist.

The FBI also has the need to measure the ability to provide a written English summary of a non-English conversation. Frequently, this conversation involves a telephone communication that has been authorized by a magistrate as part of an ongoing criminal investigation. CAL developed the Listening Summary Translation Exam (LSTE) as part of its contract with the FBI.[2]

---

[1] Special Agent Linguists are Special Agents who are qualified to investigate crimes involving foreign languages.

[2] The LSTE presents taped Spanish language conversations as stimuli and requires the examinee to answer multiple-choice questions or to provide a written summary as a response. The LSTE provides scores on the accuracy (including adequacy) of the information in the summary and on the quality of the English expression contained in the summary.

The development and validation of the LSTE is the subject of a separate report (Stansfield, Scott & Kenyon, 1990a), and is not formally treated in this report.

The FBI also has the need to measure the ability to translate written documents. Up until now, this need has been satisfied for some 20 languages through two parallel translation exams. Since these exams are secure instruments, CAL staff know nothing about them other than the fact that the FBI feels a need to develop new translation exams. Because of this, the FBI issued a request for proposals (RFP) to develop a completely new test of translation skills, which is the subject of this report and a companion report (Stansfield, Scott & Kenyon, 1990b).

1.2. Intended Use

The SEVTE is designed for use in the hiring of Language Specialists and Contract Linguists. Language Specialists are full time regular employees of the FBI, while Contract Linguists are self-employed and work on an hourly basis. The translating work of Language Specialists and Contract Linguists is primarily document-to-document or audio-to-document. The subject matter may be in any area in which the FBI has jurisdiction. As indicated on an FBI job announcement, an FBI Language Specialist is a full time employee whose duties are to "translate both recorded and written material, into English and vice versa, which involve a wide range of difficult subject matter containing technical or specialized terminology such as used in fields of law, politics, science, economics, and international exchange, as well as nontechnical subject matter."

The SEVTE would be taken by civilians who are applying for these two categories of position, and by current FBI employees, such as support staff, who are seeking a promotion to the position of Language Specialist.

According to the statement of work in the RFP, CAL is to provide a test that can measure translation ability at levels 2+ through 5. Such levels would be appropriate for Language Specialists and Contract Linguists. SEVTE scores will provide supervisors with an indication of their suitability for a given work assignment involving Spanish to English translation.

1.3. FBI Translation Needs Survey

One of the first tasks undertaken during this project was the development of a questionnaire for the purpose of conducting a survey of the type of translation work required of Language Specialists in FBI field offices. It was hoped that this survey of the FBI's translation needs would be of help in determining an appropriate balance of topics and tasks for the tests to be developed. This questionnaire was developed by CAL staff during August 1988, and was subsequently revised by the FBI. Following these revisions, FBI Headquarters mailed two copies of the questionnaire to Language Specialists working in FBI field offices across the country. A total of 28 Language Specialists replied to the questionnaire. The questionnaire concerned translating from Spanish to English and from English to Spanish. The last page of the questionnaire was devoted to translating from English to Spanish. A copy of the questionnaire and the results are included in Appendix Q. The questionnaire required the Language Specialists to indicate the proportion of time they spend translating each type of document listed in the questionnaire. Unfortunately, the results of the questionnaire are limited, since, many individual's responses totaled more than 100%. Still, the results of the questionnaire did provide supporting information for the development of the LSTE, the SEVTE, and the ESVTE. In general, the results indicated that Language Specialists spend more time doing listening tasks than translating written texts, particularly monitoring and translating telephone and recorded conversations. They are also called upon to provide oral interpretations.

More than half of the Language Specialists responding indicated they are often called upon to translate or summarize written material. The material these respondents most often deal with involves organized crime, narcotics, terrorism, and counterintelligence.

The results of this survey were used to select topics for the written and recorded stimuli that appear on the three tests developed for this project.

1.4. FBI\CAL Translation Skill Level Descriptions

1.4.1. History

Over the years there have been a number of attempts by government agencies to develop skill level descriptions (SLD) for translation. None of these have been accepted outside of the agency in which they were developed. The FBI also developed a set of translation SLDs a number of years ago. However, the Bureau was not satisfied with them. As a result, the Statement of Work in the FBI's Request for Proposals called for the development of new translation skill level descriptions (see Appendix R.) The statement of work also called for scores on the test to be convertible to the 0-5 ILR scale. As a result, CAL proposed to develop such skill level descriptions as part of this project. Once the project was funded, the first deliverable to be developed was the translation SLDs. These were needed to inform the test development process, and, in particular, to inform the scoring of the test and the conversion of the scores to the 0-5 scale. Thus, soon after notification of funding was received, CAL staff went to work on the skill level descriptions.

In July 1988, CAL staff met with the project monitor and five FBI staff at FBI headquarters. Attending were FBI master translators.[3] At this meeting it was agreed that, in order to help CAL begin the development of ILR skill level descriptions for translation, by the end of the month the FBI staff present would write a personal definition of what constitutes an excellent translator, a good translator, a mediocre translator, a poor translator, and a bad translator. It was agreed that CAL would use the descriptions of these five groups of translators as a point of departure for preparing skill level descriptions for translation. Because FBI staff were familiar with the ILR SLDs, their descriptions showed a similarity in form to these descriptions. The following description of a "mediocre" translator illustrates the kind of descriptions that were received.

---

[3] Language Specialists at FBI Headquarters in Washington, D.C. are referred to as Master Translators.

"Able to provide an understandable and fairly accurate translation of a larger number of texts, but still makes a number of mistranslations. Problems with spelling, grammar, and punctuation. Becomes lost when structure becomes complex or language more sophisticated and has serious problems with slang, idioms and handwritten materials."

The descriptions of different groups of translators provided by FBI staff, although brief and informal, were used as a starting point for writing skill level descriptions.

CAL staff began by writing descriptions for level 5 translation, and then worked down the scale to level 0+. The first set of skill level descriptions was drafted by Ana Maria Velasco, an experienced translator familiar with the ILR scale. She drafted the descriptions based on her experience evaluating the work of many different translators. In consultation with the project director, Ms. Velasco selected seven variables that should enter into the judgement or rating of a translation. These were accuracy, grammar (morphology), syntax (word order), style, tone, spelling, and punctuation. She placed these variables on the vertical axis of a scoring grid (matrix). The horizontal axis contained 10 points on the ILR scale ranging from

0+ to 5. In each cell of the grid, she included a statement of the nature of translations at that level. Both skill level descriptions and a scoring grid were developed, since it was thought that a scoring grid that separated each translation variable by level and allows comparisons by variable across levels, would be helpful to raters. It was also recognized that the grid would be useful in the revision of the skill level descriptions for the same reasons. That is, the description of ability on each relevant variable in the scoring grid could be consulted in the writing of the skill level descriptions. The final reason for producing the scoring grid was because we were unaware at the time which document, the grid or the skill level descriptions, could be used to score the test more reliably.

The project director then reviewed the skill level descriptions and the scoring grid, making revisions where appropriate. His revisions were based on careful analysis of the wording of all the current ILR skill level descriptions, particularly the reading level descriptions. The revised SLDs and the scoring grid were then subject to careful review by Marijke Walker and her staff at the FBI. They responded to the draft descriptions based on their experience evaluating the translations of Language Specialists and applicants for employment as a Language Specialist. After receiving a set of comments from Ms. Walker, CAL revised both documents. A major revision to occur at this point, at the suggestion of Ms. Walker, was the inclusion of syntax within grammar on the scoring grid and the addition of vocabulary to the grid. (A copy of the grid is included in Appendix I as Exhibit A.) Another substantive revision was a change in the percentage correct criteria for punctuation and spelling at level 5. It was decided that for purposes of the grid, the translation need not be absolutely perfect in spelling in order to be at level 5. A brief description of the kinds of documents that can typically be handled by a translator at each level was included.

On December 5, 1988, a meeting was held at FBI Headquarters to review the revised set of translation SLDs. Present at the meeting were Charles W. Stansfield and Ana Maria Velasco from CAL, Marijke Walker and her staff, Thomas Parry from the Central Intelligence Agency, and James Child from the Department of Defense. During this meeting it was noted that the draft translation SLDs describe the characteristics of the translated document, while ILR SLDs for other modes of communication describe the skills of the person being evaluated. It was suggested that the Translation SLDs should consistently describe the translator, rather than the translated document. It was also agreed to introduce this current draft of the descriptions to the ILR Testing Committee before making any revisions, and to ask committee members for written comments regarding how the draft can be improved.

These translation SLDs were the subject of a brief discussion at the December meeting of the ILR Testing Committee two days later. Members of the committee were given a questionnaire concerning the SLDs to complete and mail to CAL (see Appendix I, Exhibit B). Unfortunately, no questionnaires were returned. The committee met again in February, 1989, with essentially the same outcome. While general and conceptual concerns were expressed at the meeting about the SLDs, only three specific suggestions for improvement were made. These suggestions were a.) to change the descriptions so that they referred to the translator rather than to the translation, as suggested earlier, b.) to use the term "to render" when referring to the act of translating, and c.) to reorder the descriptions so that they begin with level 0 and progress to level 5.

Following this meeting, Charles Stansfield and Marijke Walker worked jointly on several occasions to improve the SLDs. The ILR Testing Committee met again on March 8, 1989, to consider the next revision. At this meeting it was not possible to obtain organized and coherent feedback or approval of the descriptions. Thus, CAL and the FBI agreed subsequently that the level descriptions being developed for this project would be used by the FBI, and that they would be available to the ILR for use as interim SLDs until such time as the ILR Testing Committee has time to consider and revise them further. Subsequently, Stansfield and Walker met again to make additional revisions on the SLDs. These revisions included the incorporation of some of the wording used in the previous set of translation SLDs used by the FBI. The task of developing and revising the translation SLDs was completed in June, 1989. No further work was done on them for seven months.

The Verbatim Translation Exams that CAL developed for the FBI were administered during the months of November and December 1989. After scoring the Listening Summary Translation Exam, CAL staff and consultants then scored the production portions of the verbatim translation exams. Soon it became apparent that there were limitations in the ability of the SLDs to describe all examinees. The problem seemed to lie in the fact that some examinees were translating into their native language and some into a second language. In the case of a number of examinees, there was a considerable discrepancy in the proficiency in the two languages. Examinees who were translating into their native language, especially English, produced translations that were very fluent and grammatical, but inaccurate in terms of content. Similarly, when translating into the second language, some examinees produced accurate translations that evidenced problems with grammar or vocabulary. As a result, on January 30, 1990, Stansfield and Scott sent a memo to Marijke Walker at the FBI in which they recommended that the current SLDs be divided into two parts: one for Accuracy and one for Expression, and that separate scores be assigned for each. CAL also recommended that the discussion of the kinds of documents a translator at a given proficiency level can handle be deleted from the SLDs, since the verbatim exams did not provide the opportunity to examinees to translate all of the types of documents mentioned. The FBI agreed to this change. It is most significant that the results of the validation study supported this division of translation abilities.

The current version of the SLDs is basically the same as the one that was used to score the verbatim translation exams. However, after the scoring of the test was completed, we realized that the discussion of the kinds of documents a translator at a given proficiency level can seccessfully render is useful interpretive information for test score users.[4] Therefore, the version of the SLDs included in this report, presents this discussion following the SLDs for Accuracy and Expression. It should be remembered however, that the raters of the SEVTE did not use this interpretive information when scoring the responses of examinees who participated in the validation study.

1.4.2. Explanation of the Skill Level Descriptions

The FBI\CAL translation SLDs are divided into three parts. The first part is the Accuracy description. Accuracy is the ability to correctly convey the information in the source document. The second part of the description is the Expression description. This describes the examinee's command of the written form of the target language. The third part of the translation skill level descriptions is the interpretive information. This is a sentence describing the general ability level of the examinee and the types of documents that he or she can be expected to translate successfully.

---

[4] It should be pointed out that there is no empirical data, in the form of a criterion-related or predictive validity study, to support this interpretative information.

Because an examinee may be called on to translate into his or her native language or second language, it was necessary to separate the ratings for Accuracy and Expression. By evaluating Accuracy and Expression separately, the level descriptions can be used to characterize an examinee whose translation is accurate but may evidence some problems with grammar or vocabulary. Otherwise, two different examinees might receive the same score by a rater who is attempting to compensate for either lack of Accuracy in the information conveyed or lack of grammaticality in the translation. A personnel administrator trying to make a decision on hiring would not have sufficient information from a score combining Accuracy and Expression to make an informed decision. This is because a typical profile of a level 2 (Accuracy) translator when translating into his or her native language, may be a level 4 in Expression but only a level 2 in Accuracy. Such an individual could not handle the kind of documents mentioned in the ILR reading descriptions for Level 3 or those mentioned in the interpretive information for level 3 of the translation SLDs. On the other hand, with separate scores available for Accuracy and Expression, an administrator would be able to make a decision to hire an examinee whose translations would be accurate though unpolished.

The three parts of the translation SLDs, unlike the SLDs for listening, speaking, reading and writing, must be in separate sections. This is because translation involves two languages, and the examinee's ability in each language may not be equal.

The first part of the SLDs is the Accuracy description. The Accuracy description focuses on whether the information contained in the source document is distorted or lost in the translation, or whether information has been inserted in the translation that was not in the source document. In the field of translation, such problems are referred to as mistranslation, omission, or addition. Scoring a translation for Accuracy requires comparing it with the original. The Accuracy descriptions refer to the ability to sustain performance (to render the document into the target language successfully) over a wide variety of documents varying in type and difficulty, rather than a single document. In general, Accuracy is the principal ability being measured in a test of translation. Thus, the Accuracy rating is the principal rating of the examinee's ability to translate.

Again, it must be remembered that this rating is descriptive of the ability to translate a wide variety of documents. A level three translator may translate a level 1 document perfectly, thus making it appear to be a level 5 translation. Similarly, the same translator given a level 5 document may produce a translation that appears to be less than level 3.

Because the accuracy of a translation may vary according to the difficulty of the document being translated, the developer of translation skill levels faces a dilemma. It is necessary to choose a type of document or level of document (in terms of difficulty and complexity) on which to base the Accuracy descriptions. In this case, we chose to describe Accuracy in rendering a hypothetical "average" or typical document. An average document encountered by an FBI Language Specialist, in terms of difficulty, would be one at level 3 or mostly at level 3, which would make it a 2+. As the translator moves above level 3 in ability, he or she, by definition, can handle documents of above average difficulty. That is, he or she can handle documents at level 3+, 4, or even higher. The Accuracy description nicely represents both the translation ability level of the examinee and the level of task or document that the examinee can handle adequately.

The second part of the skill level descriptions is the Expression description. Expression involves all the linguistic variables apparent in a translated document except Accuracy. These variables are grammar, syntax, vocabulary, style, tone, spelling, and punctuation. In general, it is possible to score a translation for most of these variables without referring to the source document. However, it will sometimes be necessary, especially in the case of higher level documents, to compare the source document with the translated document, particularly if the style and tone of the translated document are to be evaluated.

The discussion of the type of documents a person can handle that initiates each SLD for the other skills is not truly part of the translation scale. It is merely score interpretation information that is of interest to score users.[5]

When using the interpretive information, a score user should remember that it refers to the type of documents that an examinee can handle successfully. Efforts to translate more sophisticated documents than those associated with that level or lower levels, will result in less than adequate translations.

---

[5] If the information on the type of documents a translator can handle were to be incorporated into the translation SLDs, then a rater would have to administer the documents mentioned to an examinee in order to verify that the statement is correct. This would require some type of tailored face-to-face testing. That is, the test administrator would have to select and administer a document to the examinee. Then, the test administrator would have to wait for the examinee to render a written translation of the document. Once the rater received the document, it would have to be scored immediately. Then, the test administrator would have to select another document, associated with a higher or lower level on the scale, and administer it to the examinee, and continue the process again until the rater was satisfied that he or she had identified the highest level of document that the examinee is able to translate faithfully. To do this, would require a full day to test each examinee, which is impractical for reasons of cost. Thus, the interpretive information in the translation SLDs is not of interest to raters of translated documents.

Another theoretical possibility involving tailored testing would be to let a computer select, administer, and score the translation using the skill level descriptions as a basis for scoring. While a computer could select a document of predetermined difficulty, and administer it to the examinee, and the examinee could key-enter a translation of the document on the computer screen, it is not yet feasible for a computer to score a translation using even an analytic scale, and it is doubtful that a computer will be able to use a holistic scale (such as the SLDs) for many years to come. Thus, it is not possible to develop a tailored test of translation ability at this time. Other ILR SLDs, such as those for speaking and reading, assume that tailored face-to-face testing is possible. Thus, the inclusion in the other ILR SLDs of the type of documents or tasks that can be handled is more logical. It is not logical to include them as an integral part of the Translation SLDs.

1.5. The Nature of Translation Ability

1.5.1. The Need to Define the Construct

Bachman (1990, p. 251), citing Upshur, distinguishes between viewing a test score as a pragmatic ascription (the individual is able to perform a task), versus viewing a test score as a measure of some human construct (the individual has a certain ability). Bachman notes that there is often confusion between the measurement of the activity and the measurement of the construct and the processes that underlie it. Indeed, he notes that the activity is often confused with the construct and vice versa.

Bachman's characterization of this confusion regarding validity is somewhat analagous to the dilemma we encountered when we wrote our proposal to do this project in September 1987. In this case, we started with products (translations), and in the process of developing the test, we identified the constructs involved in the measurement of translation ability. We learned that translation ability is most appropriately expressed through two main constructs, accuracy and expression.

It is important to distinguish between translation ability as a measurement construct and translation ability as a psychological construct. A measurement construct is one that holds up under statistical analysis, such as factor analysis or other appropriate procedures. It should be supported by descriptions of the psychological construct, which refers to the mental operations and processes involved. Neither the measurement construct nor the psychological construct was understood at the start of this study. Thus, we entered the study fully aware that we were sailing uncharted waters. While hopeful that we would make some discoveries, we were fully aware that any test we constructed might not stand up to scientific analysis. Thus, we were aware that we might fail in our effort to construct a reliable and valid test of translation ability.

In terms of a psychological construct, we identify translation ability as a nexus of psychological and linguistic knowledge, skills and abilities that can be combined with real world knowledge to produce a translated document. This is an initial definition of translation as a process; it is in no sense a description of the process. At present, there is almost no understanding of the translation process. Moreover, the level of ignorance about translation is exacerbated by the fact that many translators have written about it and their writings create the impression that a literature on the process exists and, therefore, that the process is at least partly understood.

1.5.2. The Literature on Translation

The writing of translators about translation has focused on the best approach to translation.[6] Two main approaches have characterized the discussion. These are literal translation and free translation. Those who espouse a literal translation strive to be faithful to the language of the source document, while those who espouse a free translation strive to produce a similar rhetorical effect as the source document. Thus, it can be seen that academic discussions of translation center on the subject of equivalence. That is, how does one produce a target document that is equivalent to the source document.[7]

---

[6] Because the literature on translation was largely unhelpful and did not inform this test, we have not attempted to include a formal review of the literature here. Instead, we will give only a brief summary.

A discussion of this nature is far from a scientific discussion. Indeed, almost everyone who writes about translation appears to be unaware that translation is an ability that can be the subject of scientific inquiry. Moreover, when the possibility of developing a scientific knowledge base about translation is raised, it is quickly dismissed. In regards to this possibility, Newmark, who is probably the best known of those who write about translation, has stated: "There is no such thing as a science of translation, and there never will be" (1981, p. 113).

Apart from the questions of approach and equivalence, there is also some literature on the nature of a good translation, which might appear to be relevant to the measurement of translation ability. In a portion of this literature, translators usually describe some problems they encountered in translating specific documents. Another portion of this literature discusses the characteristics of a good translator or translation. The characteristics are usually stated in the form of ascriptions, i.e., is sensitive to the nuances of words in both languages, is sensitive to style, tone and purpose. Such ascriptions do not help us to understand translation as a psycholinguistic process or even the appropriate constructs to measure.

---

[7] Recently, there has been some attention to the role of text characteristics in determining the approach to use. For a summary of the rhetoric on equivalence and on the role of text characteristics, see Pochhacker (1989).

Some authors have noted that there are certain prerequisites to being a translator. Apart from the attitudinal characteristics, such as a love of language, most notable among these are a knowledge of the language of the source document, a knowledge of the language on the target document, and some knowledge of the subject.[8] Again, this information, while accurate, was not helpful to us in developing a test of translation ability.[9]

1.5.3. The Emergence of the Constructs

In this study, we identified Accuracy and expression as the measurement constructs of relevance. We define Accuracy as the ability to render the information or propositions in the source document into the target document without mistranslations, additions, or deletions. We define Expression as the ability to express oneself appropriately in the target language in the context of a translation.

---

[8] Knowledge of the subject is viewed as being less important, since it is considered that one can learn this quite easily by reading on the subject prior to beginning the translation. It is interesting to note that we did not encounter a single mention of "schema theory" in writings on translation.

[9] At the start of the study, we did a computer assited search of the ERIC database, using "translation" and "language testing" as major descriptors. The seven titles this search produced dealt with translation as a method for testing language proficiency or achievement. Not a single one dealt with the measurement of translation ability per se.

We could not identify these constructs at the start of the project. Instead, they emerged slowly as the project progressed. As indicated in section 1.4., the first task in this project was the development of skill level descriptions (SLDs). These SLDs combined statements referring to Accuracy, to categories of expression, and to the type of documents a translator can handle. The SLDs were written so that they could be used in some way when scoring the test or referenced when interpreting the test score. Once the descriptions were drafted, we began developing the tests.

The process of scoring trial tests and pilot tests provided us with more experience in the measurement of translation. For instance, pilot testing taught us that people performed much better when translating into their native language. Thus, we learned that a single set of skill level descriptions could not be used to characterize translation ability in both directions. For the sake of parsimony, we had initially hoped that it would be possible to characterize a translator through a single proficiency rating that would indicate his or her ability to translate in both directions; that is, from native language to target language and from target language to native language. While this may seem naive in retrospect, at the time we were influenced by the elimination of the distinction between native languages and second languages in linguistics (see Kachru, 1985), since proficiency in either can range from almost none to distinguished. Thus, we were not willing to accept the recommendation that separate sets of SLDs be developed for translating in each direction. Since we believed a single set of SLDs would be adequate, we also believed that a single rating could characterize translation ability in both directions, and that separate ratings for each direction were not necessary. The experience of scoring pilot tests which were given in both directions made us doubt this assumption and in the ensuing months we abandoned the idea entirely. Still, we believed, and we continue to believe, that the same set of SLDs can be used for both directions, and that the development of a separate set of SLDs for translating to the native language and another for translating to the second language is unwise.[10] Thus, we began the project believing that a single holistic score could represent translation ability, and by the end of the pilot testing we had modified our ideas so that we now believed that two scores, one for translating in each direction, would be necessary.

At this point another experience began to influence our ideas. During the fall of 1989, we administered, scored, and analyzed the Listening Summary Translation Exam. This test, which is the subject of another report (Stansfield et al., 1990a), produced two scores, one for Accuracy and one for

---

[10] A number of government translators advised us to do this.

Expression. A separate score for Expression had always been considered for this test, since we were aware that errors in English writing ability have posed a problem for the FBI when translations oral conversations are introduced in court. That is, even if a translation is accurate, if it is written poorly, the credibility of the information it contains becomes tainted.

The analysis of the LSTE showed the validity of the Accuracy rating in terms of its correlation with other measures of proficiency in the language of the auditory stimuli. The analysis also showed Expression to be an entity different from and often unrelated to Accuracy. As a result, we concluded that Accuracy is the principal trait to be measured in a test of listening summary writing ability, but that it may also be useful to have an expression score in order to identify examinees whose work may need to be reviewed before being used in a legal proceeding.

As indicated in section 1.4.1., soon after scoring the LSTE, we began scoring the SEVTE and a parallel test in the opposite direction, the English - Spanish Verbatim Translation Exam (ESVTE). We soon realized that it would not be possible to use the SLDs to score the paragraph translation portion of these tests since the performance on the criteria relating Accuracy was often incongruous with the performance on the criteria relating to Expression. At that point, it became apparent that the solution to this problem lay in considering Accuracy and Expression as separate constructs and assigning separate scores to each. This decision to divide translation ability into two constructs is supported by the many analyses reported in the section on validity of this report. Thus, while we began this project believing that translation ability in both directions could possibly be represented in a single rating, we ended the project having learned that four scores are necessary to represent translation ability, i.e., two for each direction. These scores do not describe the psychological construct or ability, but they do identify and define the measurement constructs.

In order to gain an understanding of the psychological construct, psychologists and applied linguists will have to turn their attention to the process of translation. A description of these processes is essential to understanding the construct of translation ability.

Due to the lack of relevant research on translation, this project was begun without an understanding of the construct to be measured. We ended the project without an understanding of the process of translation, but with the belief that we at least subdivided the construct in a practical way so that instruments can be developed to measure it. We believe the instrument described in the remaining sections of this report is a good one. However, in the coming decades other researchers will develop other instruments that may have greater reliability, due to improved scoring procedures, or greater validity, due to a better understanding of the psycholinguistic processes involved in translation. Nevertheless, it is likely that high quality instruments to measure translation ability will continue to focus on the constructs of Accuracy and Expression that emerged from this project. Thus, at this point, for the purpose of measurement, we believe it is possible to define the construct of translation as the ability to render accurately content information from a source language text to a target language text and the ability to express this information using appropriate target language grammar, syntax, vocabulary, mechanics, style, and tone.

2. General Description

The Spanish into English Verbatim Translation Exam (SEVTE) is designed to assess the ability to render a verbatim translation in English of source material written in Spanish.

The SEVTE consists of two subtests. The first, referred to in this part of the report as the Multiple Choice section, consists of embedded phrase translation and error detection items. The second subtest, referred to as the Production section, requires translation of embedded phrases, sentences, and paragraphs. A separate test booklet, containing instructions, examples, and test items, is provided for each subtest. There are two forms of the SEVTE; they are generally parallel in content, item difficulty, format, and length.

2.1. Multiple Choice Section

This section of the report describes the format, and test taking and scoring procedures for the Multiple Choice section of the SEVTE.

2.1.1. Format

There are 60 items in the Multiple Choice section: 35 are Words and Phrases in Context (WPC) items, and 25 are Error Detection (ED) items. In a WPC item, an examinee is required to select the best translation of an underlined word or phrase within a sentence. In an ED item, an examinee must identify where an error is located within the sentence, or indicate that there is no error. ED items are written in the target language only; errors may consist of incorrect grammar, word order, vocabulary, punctuation, or spelling. (There is no more than one error per item.)

The multiple choice items are designed to test specific grammar points such as subject-verb agreement, verb tense (preterit vs. imperfect, subjunctive, etc.), pronouns, prepositions, gender, or word order; or vocabulary, including noun, verb, adverbial, and adjectival phrases, and false cognates. The results of a content analysis[11] of the SEVTE Multiple Choice sections are displayed in Appendix D. Briefly, 30-32% of the items assess knowledge of grammar, 60% assess knowledge of vocabulary, 8% assess knowledge of mechanics (spelling or punctuation), while 5% of the items contain no error.[12]

The test booklet contains instructions, example items for each subsection (WPS and ED), explanations of the example items, and the test items. Appendix B contains selected portions of a test booklet for the Multiple Choice section, including the cover page, instructions, and example items. This appendix can be used by the FBI to construct an examinee handbook.

2.1.2. Test Taking

Each examinee receives a Multiple Choice section test booklet, a machine scoreable answer sheet, and two no. 2 pencils.

---

[11] The content analysis of test was carried out by CAL staff and then verified by FBI Headquarters staff.

[12] Some of the items test knowledge of more than one aspect of language.

Examinees listen as the test supervisor gives instructions for filling out the machine scoreable answer sheet and the test booklet cover page. Subsequently, they are given 35 minutes to complete the Multiple Choice section.

2.1.3. Scoring Procedures

Examinees record their responses to the Multiple Choice section of the SEVTE on answer sheets which are scored by machine. The score on this section is the number of answers correct. The maximum possible score is 60.

2.2. Production Section

This section of the report describes the format of the Production section as well as test taking and scoring procedures.

2.2.1. Format

There are 28 production items on each exam form; 15 items, called Word or Phrase Translation (WPT), require translation of underlined words or phrases in sentences, 10 items, called Sentence Translation (ST), require translation of complete sentences, and three items, called Paragraph Translation (PT), require translation of entire paragraphs.[11]

The test booklet contains instructions, an example of each item type (except for the paragraphs), a brief discussion of each example item, and the test items. Space is provided in the booklet for the examinee to write the translation below each item. Appendix C contains selected portions of a test booklet for the Production section, including the cover page, instructions, and example items. The reader may find it helpful to refer to these now in order to get a better understanding of the nature of the SEVTE.

---

[11] The paragraphs on the SEVTE forms range from 87 to 121 words in length, averaging 99 words per paragraph.

2.2.2. Test Taking

Examinees are given 35 minutes to complete the first two subsections (WPT and ST) and 48 minutes to complete the paragraph subsection. They are permitted to use dictionaries only in translating the paragraphs.

2.2.3. Scoring

As noted above, examinees write their translations in the test booklet. Each subsection is scored by a trained rater according to the procedures outlined below.

2.2.3.1. Words or Phrases in Sentences Items

The keys for this subsection are quite comprehensive, containing a number of acceptable translations for each item. However, when scoring the test a rater is free chose to accept other appropriate translations that are not included in the key if he or she believes that translation is correct. The items are scored as either correct or incorrect, regardless of whether an error consists of incorrect grammar, word choice, or syntax. One point is awarded for each correct translation; hence, the maximum score for this subsection is 15 points.

2.2.3.2. Sentence Translation Items

The keys for this subsection contain several acceptable translations for each item, although the keys do not purport to list all possible acceptable translations. A trained rater assesses the Accuracy of the translations, i.e., the extent to which the original meaning has been appropriately conveyed. From 0 to 5 points are awarded for the translation of each sentence, according to the scoring guidelines found in Appendix E. As there are 10 sentences, a maximum of 50 points are possible for this subsection.

2.2.3.3. Paragraph Translation Items

The keys for this subsection provide only one translation for each paragraph, even though a number of slightly different but acceptable versions are possible. The example translation is intended to provide a standard interpretation of the source text, and raters may use their expertise in the language to judge whether variations in examinee renditions remain faithful to the original meaning. On the other hand, the rater training materials provide several examples of translations at different ability levels, along with appropriate scores for each translation.

Examinee translations are evaluated for correctness of Grammar (morphology), Expression[14] (in the case of the paragraph translation items only, Expression refers to word order and vocabulary), Mechanics (spelling and punctuation), and Accuracy (as described above). From 0 - 5 points are awarded in each category according to the guidelines located in Appendix F. Since there are three Paragraph Translation items, a total of 60 points are possible for this subsection; 15 points for Accuracy and 45 for Expression.

---

[14] The reader is advised not to confuse paragraph expression with the overall Expression score. The overall Expression score includes all criteria referred to in the SLDs other than Accuracy.

2.3. Computation of Total Scores

A total score is computed separately for Accuracy and Expression. (See the discussion of these constructs in section 1.5.3) A maximum score of 185 points (80 for Accuracy and 105 for Expression) is possible for the entire exam. The total for Accuracy and Expression is then converted to a Translation proficiency rating (one of the new CAL/FBI Skill Level Descriptions) using the conversion tables (one for each exam form) found in Appendix O. The development of these conversion tables is described in section 6.3 of this report.

The total score for Expression is composed of the 60 items in the Multiple Choice section, which are worth up to 60 points, plus the sum of the points earned for Grammar, Expression, and Mechanics (up to 45 possible) on the Paragraph Translation subsection of the Production section. Thus, the examinee may obtain a raw score of up to 105 points for Expression.

The total score for Accuracy is composed of the 80 points that may be earned on the Production section. The examinee may earn 15 points for Accuracy in the Word and Phrase Translation items, 50 points for Accuracy in the Sentence Translation items (up to 5 points for each of 10 sentences) and 15 points for Accuracy on the three paragraphs (up to five points per paragraph).[15]

2.4. Use of Multiple Choice Section for Screening

The Multiple Choice section may be used to screen out individuals for whom the Production section of the exam would be inappropriate. Since the minimum recommended passing score is 2.8 or a 2+ on the Translation Skill Level Descriptions, examinees should not be screened out who have some reasonable chance at scoring at this level. Prior FBI policy has established a 2.0 as a screen (previously based on a DLPT reading score), and CAL was requested to continue this practice by using the Multiple Choice section score corresponding to a 2.0 on the entire SEVTE as a screen. Through statistical analyses (described in section 8.4), we have determined that the raw score cut-off on the Multiple Choice section should be 33 for Form 1 and 25 for Form 2. Examinees scoring at or below these scores need not take the Production section of the SEVTE, since they are unlikely to have a translation skill level at 2.8 or above when the entire exam is administered. If they have already taken the Production section, it need not be scored.

---

[15] As explained later in this report, a multiple regression analysis did not improve on this raw score weighting. Thus, it was decided to use this weighting to calculate the total score for Accuracy. The effect of this weighting is that the Sentence Translation subsection counts more than three times as much as the Paragraphs subsection.

3. Development of the SEVTE

This section describes the development of the two pilot forms of the SEVTE. The preparation of examination materials and the development of pilot study scoring methods are also discussed.

3.1. Exam Forms

Items for the SEVTE were developed by CAL staff and consultants, taking into account the results of the survey of FBI translation needs (see section 1.3), the results of which are reported in Appendix Q of this report. They relied on their expertise as translators and teachers in developing the items. The item developers sought to test aspects of Spanish that are especially challenging to translate because there is no direct equivalent in English. The developers also focused on aspects of grammar that have traditionally caused problems for Spanish/English translators and students because there is no direct correspondence between the two languages. These areas include pronouns, verb tenses and sequence of verb tenses, use of negatives, possessives, prepositions, and non-temporal verb forms (infinitives, gerunds, past participles), among others.

A number of item texts were either excerpted directly from documents provided by the FBI or were paraphrases of such documents. In addition, many items were paraphrased from newspaper and magazine articles and documents encountered in the professional work of the item developers. The developers selected the material carefully, so that the topics and vocabulary of the item texts would be consistent with the type of documents FBI employees reported being required to translate on the survey of FBI translation needs.

Parallel forms were organized by matching items according to point being tested (specific grammar point or vocabulary) and by matching them in terms of difficulty on the FBI/CAL SLDs for translation. This latter matching required the test developers to make an estimate of the difficulty of rendering the translation, rather than of the difficulty of the language of the item itself in either the source or target language. The items were originally arranged in order of increasing difficulty. More items were developed than we anticipated would be needed on the final forms, so that items that did not function effectively could be discarded after pilot testing. Originally, there were 63 items (35 Words or Phrases in Context and 28 Error Detection) in the Multiple Choice section of Form 1, and 64 items (35 Words or Phrases in Context and 29 Error Detection) in the the Multiple Choice section of Form 2. The Production sections of both forms contained 23 Word or Phrase Translation items, 16 Sentence Translation items, and three Paragraph Translation items.

Following extensive internal review, CAL sent the SEVTE exam forms to the FBI for preliminary approval and revised them according to FBI suggestions prior to trialing.

3.2. Pilot Test Scoring Procedures

Answer keys were prepared for the Multiple Choice and Production sections. The keys were reviewed by FBI staff members, and a number of their suggestions were incorporated in making revisions.

Examinee responses to the Multiple Choice section were to be scored by an optical scanner, which would tabulate the number of correct answers. Similarly, examinee translations of the Word or Phrase Translation items in the Production section were to be scored by raters as being either correct or incorrect, according to the keys which had been prepared.

In contrast, scoring of the Sentence Translations and Paragraph Translations was to be based on the new FBI/CAL Translation Skill Level Descriptions. The Translation Skill Level Descriptions were intended to characterize an examinee's performance on a range of materials. Thus, it was not possible to use them to score individual sentence items because these item texts were too restricted. Consequently, CAL staff developed simplified scoring guidelines, based on the FBI/CAL translation skill level descriptions, for evaluating both ST and PT items.

In preparation for writing the simplified guidelines, the FBI/CAL skill level descriptions were reorganized so that all proficiency levels were described within each category, i.e. Grammar, Syntax, Vocabulary, Mechanics, Accuracy, and Style and Tone. (For example, references to grammar in levels 0+ - 5 were all placed on the same page.)

After studying these reorganized skill level descriptions, an attempt was made to characterize each level succinctly within each category. The plus levels were eliminated, so that the scale consisted of 0 - 5 points in each category. Because exam texts were based primarily on legal and business documents (i.e., formal writing), which did not vary much in terms of Style and Tone, it was decided not to include Style and Tone as a separate category in the scoring system. The Vocabulary category was also eliminated, since aspects of this category could be subsumed under Expression and Accuracy. Finally, correctness in Mechanics (spelling and punctuation) was expressed in terms of numbers of errors for the Sentence Scoring Grid, and proportions of items correct for the Paragraph Scoring Grid. The pilot version of the Sentence Scoring Grid is located in Appendix G; the Paragraph Scoring Grid can be found in Appendix H.

4. Trialing and Pilot Testing

This section describes the trialing and piloting of the SEVTE. The results of the piloting and subsequent revisions are also discussed.

4.1. Trialing

The trialing of the two forms of the SEVTE was carried out at CAL on February 20 and 21, 1989. Three CAL employees and one CAL spouse took the exams. The Spanish oral proficiency levels of these four people varied from level 2+ to level 5, the latter being a practicing attorney who is an educated native speaker from Argentina.

Before taking each form, examinees received a questionnaire that asked them to provide a global rating of their English and Spanish proficiency (see Appendix J). After completing each section of the test, they commented on it and noted on the exam feedback questionnaire (see Appendix K) specific errors or problems they encountered.

CAL examined the responses to each item as well as to the questionnaire in order to determine which items should be modified and which should be deleted, and the exam forms were revised accordingly.

On March 29, 1989 two FBI translators each took either Form 1 or Form 2 of the SEVTE. They provided written feedback to CAL which was taken into consideration in revising the exams after the pilot testing.

4.2. Pilot Testing

This section describes the SEVTE pilot data collection, the results of pilot testing, and the revisions that were made following data analysis.

4.2.1. Data Collection

The SEVTE exam forms were piloted at Georgetown University on April 1, 1989. Forty-five students from the Department of Translation and Interpretation completed the Multiple Choice sections of both forms together as a group. Each student was paid $25.00 for taking both sections. Graduate students in the Translation Certificate program took the complete exam; four students took Form 1 and five took Form 2. Each of these students was paid $15 for taking one form of the entire SEVTE exam.

The Georgetown University students kept track of how many minutes it took them to complete each section of the exam. They also completed a questionnaire regarding their native language background and their proficiency in English and Spanish. (Appendix M contains a copy of the questionnaire; a summary of the responses of examinees is also located in Appendix M. The data in this summary represents all examinees who participated in the pretesting, including those graduate students who took either the SEVTE or the ESVTE.) In addition, we asked students to comment on any items that were confusing or that caused them particular difficulty.

Of the 48 students who participated in the pretesting, English was the native language of 41. 7 students indicated another native language, but knew some Spanish. These other native languages were Portuguese, Tagalog, Korean, Chinese, Russian, and Italian.

4.2.2. Results

Table 1 displays a summary of the performance of the pilot study examinees on the Multiple Choice sections of the SEVTE exam forms. Reliability estimates, calculated using Kuder-Richardson formula 20 (KR-20), are also shown.[16]

Table 1
SEVTE Multiple Choice Sections
Total Pilot Sample

Form	N	Mean	%	Std. Dev.	KR-20
1	47	45.6	72	5.65	.73
2	48	48.0	75	6.01	.76

There were 63 items on the pilot version of Form 1, and 64 on Form 2. Using the mean percentage correct to compare the two forms, it is apparent that Form 1 was slightly more difficult than Form 2, although both forms appeared to be somewhat easy for this group of examinees.[17] The reliability estimates were low, indicating that some of the items were not functioning as intended (i.e., they were either too easy or too difficult, or failed to discriminate among high and low proficiency examinees).

---

[16] 13KR-20 yields an estimate of the internal consistency of the test items, i.e., a measure of the extent to which examinees perform consistently across the items within a test. It is very similar to parallel form reliability.

[17] A four-option, multiple choice exam of optimal difficulty would exhibit a mean score of 62.5% correct.

A record was kept of the time it took students to complete the Multiple Choice sections. The amount of time required ranged from 24 to 31 minutes.

Since only a few examinees took the Production sections, descriptive statistics for this section were not calculated. The principal goals in piloting the Production sections were to evaluate the appropriateness of the scoring system, and to identify items that were either ambiguous, too easy, or too difficult.

4.2.3. Revisions

Students were divided by native language background (English, and other), and item analyses were conducted of their responses to the Multiple Choice section items. The item analyses showed that the items were easier for the native English speakers. (A majority of those who participated in the piloting were native English speakers.)

Seven nonnative speakers of English, from backgrounds other than Spanish, also took the SEVTE. (Unfortunately, no native Spanish speakers took this exam.) Since the item analyses showed that many items on both forms of the Multiple Choice section were quite easy for nonnative as well as for native English speakers, it was necessary to write a number of new items and to revise many of the existing items to make them more difficult. The revision process involved deleting some items entirely and replacing others with new items that assessed a similar grammar point or vocabulary item. Some of the distractors in a number of the remaining items were also modified. In addition, items that did not discriminate well among high and low proficiency examinees in the total sample were eliminated. Finally, comments written by students after completing the exam were taken into consideration in identifying items for revision. We decided to include 35 Word or Phrase in Context items and 25 Error Detection items, for a total of 60 items, in the final form of the Multiple Choice section. This is slightly fewer than the 63 and 64 items included on the field test versions of the SEVTE.

For the final version of Form 1, 30 (50%) new items were developed, and 23 (16%) distractors were modified; for Form 2, 27 (45%) new items were developed, and 20 (14%) distractors were revised. In general, the new items were designed to be more difficult, while the distractors were rewritten so that they would be more attractive to examinees.

Responses to the Production sections were scored by CAL staff and consultants in order to try out the scoring procedures and to gather information that could be used in revising items. As with the Multiple Choice section, the Production section items were analyzed in light of student performance (and comments from FBI staff as noted above). It was decided to include 15 embedded phrase, 10 sentence, and 3 paragraph translation items on the final versions of the exam forms. Seventeen (59%) of the phrase and sentence items were deleted from Form 1, and 3 new items were created; 18 (62%) were deleted from Form 2, and 4 new items were created. None of the paragraph items were modified.

The test booklets were revised to reflect the changes described above and copies were made in preparation for the validation study described in section 5 of this report.

5. Validation Study

The purpose of the SEVTE validation study was to assess the reliability and validity of the SEVTE as a measure translation ability. In this context, the validation study had a number of specific aims. One aim was to field test the revised exam to see if its items and sections performed acceptably. Another aim was to administer the test to a more appropriate population than the pretest versions' population in order to set passing scores based on their performance.[18] Another aim was to further assess the rating criteria that had been developed for scoring each part of the Production section. Another was to determine whether this section could be scored reliably. The validation study, as the word validation implies, also sought to gather information on the validity of the test. With the analysis of construct validity in mind, it was decided to collect scores on other measures from employee files and to assess the test's ability to predict overall translation ability by having raters make an overall assessment of ability using the FBI/CAL Translation SLDs. Another aim of the validation study was to gather evidence concerning criterion-related validity by having examinees rate their ability to translate various types of texts on the job, and then determine the relationship between scores on the test and the self-ratings. We chose to use self-ratings, rather than supervisor's ratings, because we were advised by the FBI that supervisors would not be in a position to evaluate translation ability. Another aim was to determine if examinees felt the test to be a valid test of their translation ability. An additional aim was to gain a further understanding of the constructs the test measured; at the time we were not sure if we were measuring a single construct, two or more constructs, or whether we were measuring a test method effect (recognition versus production).[19] Another purpose of the validation study was to determine the most appropriate weighting of the parts and sections. A final purpose of the validation study was to gather the data necessary to equate the two parallel forms of the test. This section describes the validation study design, and data collection procedures. The results of the study are discussed in the following three sections.

---

[18] The population that took the field test version consisted mostly of university students.

5.1. Overview

The orignal design of the validation study called for administering the SEVTE to FBI Language Specialists and Contract Linguists at various field offices around the country. It was hoped that individuals of varying ability levels would be included in the sample. In order to examine the validity of the SEVTE, scores on other measures of language ability were obtained from employee files as available.

---

[19] This degree of uncertainty and the multiple aims of the validation study were due to the fact that so little was known about the measurement of translation ability at the time the project began. Thus, the validation study, and indeed the entire project, combined experimentation with a commitment to develop and validate a test. To draw an analogy to the business world, it is as if we were carrying out both the research and development function and the manufacturing function at the same time. Under normal circumstances the manufacturing function is carried out after the R+D function has been completed. While far from ideal, the reality of our situation was that we were working under a fixed-price contract to manufacture a test. The client was aware of the possibility of R+D problems, but it was assumed that these would be worked out along the way.

Both forms of the SEVTE were given in one sitting (about four hours in duration) at each of seven FBI field offices. The order of administration of the forms was counterbalanced to control for the practice effect. Thus, approximately half of the examinees took Form 1 first and the other half took Form 2 first.

5.1.1. Test Administration Instructions

CAL developed a set of test administration instructions for the SEVTE. These include instructions to the test administrator regarding the following: 1) test security, 2) assembling test materials, 3) arranging for a testing site, 4) equipment, 5) administering the test (including timing of sections), and 6) procedures to follow after the test. Appendix A contains a copy of the administration instructions for the SEVTE.

5.1.2. Questionnaires

CAL developed two questionnaires for use in the validation study: 1) a self-assessment questionnaire on which an examinee was asked to estimate his or her ability to render a verbatim translation from Spanish into English, and 2) a questionnaire requesting examinee feedback on aspects of the format and content of the exam. (A copy of the self-assessment questionnaire is located in Appendix N, and a copy of the exam feedback questionnaire is in Appendix L.)

5.1.3. Subjects

Testing materials, including test administration instructions, numbered test booklets, answer sheets, pencils, questionnaires, and test administrator report forms[20] were sent to the FBI field offices in Los Angeles, San Diego, Albuquerque, Phoenix, and El Paso on November 15, 1989. Similar sets of materials were sent to Houston[21] and Puerto Rico on November 17, 1989.[22] Materials from SEVTE administration were returned to CAL within two to eight weeks.[23]

---

[20] CAL developed this form for test administrators to note any irregularities that may occur with respect to test security, the test administration, or the condition of the test materials. We requested that the validation study test administrators complete and sign the form even if there were no irregularities. (See Appendix A for an example of this form.)

[21] Arrangments were made for members of the Houston Police Department (for whom Spanish OPI scores were available) to be tested along with the FBI employees at the Houston field office.

[22] A cover letter was sent with the materials to the contact person at each field office. In addition to thanking them for their assistance in carrying out the validation study, the letter emphasized the importance of test security, outlined the procedures for the test administration, noted the proposed administration date, and instructed them to return all materials to CAL immediately after the test administration. A checklist of the materials was enclosed with each cover letter. CAL retained a copy of the checklists and used them to verify that all of the materials were returned as requested.

[23] Although most field offices were able to follow the administration procedures as outlined, a few had difficulty scheduling all of the examinees to be present for the test administration, and consequently had to give more than one administration of the same exam. These difficulties accounted for their delay in returning some of the exam materials.

Since the FBI Language Specialists were already working in Spanish, there were no examinees with low level translation ability among them. Also, because of the dire need for the services of the FBI's current Language Specialists, it was difficult to recruit an adequate number of Language Specialists and Contract Linguists for the validation study. Thus, in an effort to ensure a minimally adequate sample size, and to ensure that the entire range of abilities of potential test takers in the operational program (the testing program for applicants) would be represented in the sample, the FBI and CAL arranged for 13 beginning Spanish language students at the CIA to take the SEVTE Multiple Choice sections during the first week of April, 1990. Also, FBI Field Offices were allowed to assign Special Agents and bilingual support staff to take the test. In addition, CAL contracted three professional translators to take the full SEVTE forms. These exams were administered at CAL on January 9, 1990.

Hence, a total of 58 examinees took the SEVTE in the validation study. Of this group, 15 (26%) were FBI Special Agents, 11 (19%) were FBI Language Specialists (or Contract Linguists, who do similar work), 10 (17%) were FBI support staff, 6 (10%) were members of the Houston Police Department, 13 (22%) were CIA Spanish language students, and 3 (5%) were professional translators. It should be reiterated that while it was originally envisioned that the subjects of the validation study would be limited to Language Specialists, we were unable to secure release time for an adequate sample of Language Specialists to take the test. After discussing alternatives with FBI Headquarters staff, it was decided to include other FBI personnel (Special Agents and support staff) in the validation sample, as well as the other groups that were represented.

5.2. Scoring

The Multiple Choice parts of the SEVTE forms were scored by machine, using answer keys based on the revised versions of the forms.

The Production parts were scored by the same raters (Matilde Farren and Mary Lee Scott) who scored the pilot study data, using the scoring keys and analytic sentence and paragraph guidelines which had been prepared. Word and Phrase Translation items were scored using a key of acceptable responses, which has been provided to the FBI. Sentence Translation items were scored using the Sentence Accuracy Scoring Guidelines (See Appendix E). These focused on the the presence of mistranslations, omissions, and inappropriate additions in the content of the translation, as well as on the conveyance of all appropriate nuances.

In order to determine which scoring system was most efficient and yielded the highest interrater reliability, the Paragraph Translations were scored in two ways, a) using the analytic paragraph guidelines, and b) using the FBI/CAL translation skill level descriptions. The SEVTE Paragraph Scoring Guidelines (see Appendix F) require the rater to assign each paragraph from 0-5 points on each of four criteria:

grammar, expression, mechanics, and accuracy. The totals for the first three criteria, grammar, expression, and mechanics, are summed to produce the Expression score for the Production section. The ratings from Accuracy are summed and contribute to the total Accuracy score, which is earned exclusively on the Production section of the SEVTE. The scoring guidelines for grammar require the rater to distinguish between errors in simple and complex structures, between low frequency and high frequency structures, and to consider the number of errors of each type in each paragraph. The scoring guidelines for expression require the rater to evaluate the paragraph for word order, vocabulary, idomaticity style and tone. After consideration of these, the rater makes a judgement as to the degree to which the translation follows the conventions of the source language or the target languages. The scoring guidelines for mechanics require the rater to evaluate each paragraph for the frequency of errors in spelling, punctuation, and capitalization. The scoring guidelines for Accuracy are identical to the scoring guidelines for Sentence Translation items. Additional information on the scoring procedures can be found in sections 2.1.3 and 2.2.3 of this report.

After the scoring of the Production section was complete, each rater assigned an overall ability level for Expression and Accuracy using the FBI/CAL SLDs, based on evaluation of the sentence and paragraph translations. This overall ability level was used in order to construct the FBI/CAL Translation Scale conversion tables.

It should be noted that initially it was hoped that a single translation ability level could be assigned to each examinee. The decision to score Expression and Accuracy separately was made by CAL after the data were collected as a result of experience gained during the pilot study and after the scoring of an initial group of SEVTE papers from the validation study. This decision was made to aid in evaluating different types of examinee performance. Some translations were very fluent and grammatical but inaccurate (as may occur when an examinee's proficiency is higher in the target language), while others were mostly accurate but evidenced problems with grammar or vocabulary (as may occur when an examinee's proficiency is higher in the source language).

In order to be able to assign separate FBI/CAL Expression and Accuracy scores, the original FBI/CAL translation SLDs were reorganized so that the descriptions for Expression at each level were contained in one section and the descriptions for Accuracy in another. A copy of the reorganized SLDs can be found in Appendix I.

6. Reliability

The data on reliability that resulted from the validation study test administration are presented in this section by subtest. An effort was made to examine reliability in a number of ways and from a number of perspectives. It should be remembered that this data on reliability is a function of the sample tested and the raters used.

6.1. Multiple Choice Section: Descriptive Statistics and Reliability

Table 2 presents the results of the validation study administration of the Multiple Choice section of the SEVTE forms. This section is referred to here as MC1 and MC2.

Table 2
Descriptive Statistics for SEVTE MC1 and MC2

Form	N	Mean	Std. Dev.	Minimum	Maximum
MC1	58	37.5	9.60	9	57
MC2	58	34.9	10.78	9	55

As can be seen in Table 2, the mean score on MC2 was 2.6 points lower than on MC1. Thus, MC2 appears to be somewhat more difficult than MC1. However, given the magnitude of the standard deviation on both tests, the difference between the two means is not significant. The larger standard deviation for MC2 suggests that less competent examinees may have tended to score slightly lower and more competent examinees slightly higher on MC2 than they did on MC1.

As there were a total of 60 items in the Multiple Choice section, the mean of MC1 represents 62.5% correct, while the mean of MC2 represents approximately 58.2% correct. Thus, MC1 appears to be of optimal difficulty, while MC2 is slightly more difficult than would be ideal for this sample.[24] Indeed, the lowest score on both forms (9) was quite a bit lower than what would be expected by chance alone (15). This apparently occurred because a few of the lower ability examinees were not able to complete the Multiple Choice section in the time allotted.

Table 3 presents the KR-20 reliability estimates for the two forms of the Multiple Choice section based on the validation study sample. KR-20 is a measure of internal consistency reliability, which is the degree to which the items (considered as a set) on a test measure the same ability.

Table 3
KR-20 Reliability for MC1 and MC2

Form	KR-20
MC1	.89
MC2	.91

The reliability of the Multiple Choice section of both SEVTE forms is high and indicates that either form can be used with confidence on a population similar to that of the validation study.

---

[24] We would expect a mean around 62.5% on a four-option, multiple choice test of optimal difficulty for the population, when the sample fully and equally represents the total range of abilities in the population.

A second indication of the reliability of the section is the consistency of performance of the group of 58 subjects on the two forms. Referred to as the coefficient of equivalence or parallel form reliability, this type of reliability is obtained by calculating the Pearson Product Moment correlation between subjects' performance on the two different forms. For the multiple choice section on the two SEVTE forms, the coefficient of equivalence is .81. This is within acceptable limits. Together, both the KR-20 reliability estimates and the coefficient of equivalence are adequately high, indicating that the two main sources of measurement error (inconsistency across items and inconsistency across forms) are minimal for the Multiple Choice section of the SEVTE.

6.2. Production Section: Descriptive Statistics and Reliability of the Accuracy Score

Table 4, which follows, shows the descriptive statistics for the SEVTE-Accuracy Subsections and Totals by form and by rater. Close examination of the means in Table 4 shows that the difficulty of the two forms is very similar. Averaging the scores assigned by both raters, we see that the Word and Phrase Translations seem to be slightly harder on Form 2 (7.0 versus 7.85 on Form 1), while the Sentence Translations seem to be slightly harder on Form 1 (28.55 versus 30.0 on Form 2). The Paragraphs seem to be equally difficult on both forms (6.65 on Form 1 and 6.55 on Form 2). The two raters appear to be consistent in their degree of severity, with Rater 1 always being more generous than Rater 2, except in the case of the Sentences on Form 2, where they are equally severe.

Table 4
Descriptive Statistics for SEVTE Accuracy
Forms 1 (N=45) and Form 2 (N=44)

Measure	Mean	Std. Dev.	Minimum	Maximum
Word + Phrase				
R1 F1	9.2	3.0	3	15
R2 F1	6.5	2.7	2	13
R1 F2	8.0	3.0	2	14
R2 F2	6.0	2.9	2	12
Sentences				
R1 F1	29.2	11.1	4	46
R2 F1	27.9	9.3	9	47
R1 F2	30.0	9.4	6	47
R2 F2	30.0	7.6	9	46
Paragraphs				
R1 F1	7.0	3.5	0	14
R2 F1	6.3	2.7	1	12
R1 F2	7.6	3.0	0	14
R2 F2	5.5	2.6	0	11
Total				
R1 F1	46.46	15.88	14.5	73
R2 F1	41.72	12.73	16	67
R1 F2	46.97	12.74	26	74
R2 F2	42.72	10.49	23	68

Legend: R=rater, F=form. Thus R1 F1 is the score assigned by rater 1 on form 1.

In discussing the reliability of the SEVTE Accuracy scores, there are two sources of measurement error that need to be examined: inconsistencies across raters and inconsistencies across forms. Traditionally these have been examined separately, though today generalizability theory allows us to look at both together. In this discussion we will first examine these two sources of error separately by examining interrater reliability and parallel form reliability. We will conclude with an examination of the results of a generalizability study on the data.

Table 5 shows the interrater reliability (Pearson Product Moment Correlations) of the SEVTE Subsections and the total Production section score for Accuracy. The reliability for Form 1 is listed first, followed by the reliability for Form 2.

Table 5
Interrater Reliability of
SEVTE Production Subsections and Production Total
for Accuracy (Forms 1+2)

	Form 1	Form 2
Word and Phrase	.86	.85
Sentences	.89	.90
Paragraph (Accuracy)	.74	.78
Total Accuracy	.93	.93

As can be seen, the interrater reliability estimates of the Accuracy scores on all subsections are quite high, with highest correlation for Sentence Translation. Across the two forms, the correlations for each Accuracy subsection are also highly similar. The interrater reliability estimates for the total Accuracy score (.93) are high and consistent across forms.

Table 6 presents the coefficient of equivalence of the Accuracy scores across forms and raters. This data is an indication of the parallel form reliability of the SEVTE across different raters.

Table 6
Coefficient of Equivalence for SEVTE Accuracy Scores
(N=43)

	Form 2 Rater 1	Form 2 Rater 2
Form 1 Rater 1	.85	.89
Form 1 Rater 2	.89	.89

As can be seen, the coefficient of equivalence of the SEVTE Accuracy score is quite high for a free response test scored by a single rater. That is, there is a high degree of agreement across forms and raters. This suggests that SEVTE Accuracy scores can be highly stable. Even under the most severe circumstances, an examinee taking different forms which are in turn scored once by a different rater, the scores show a remarkable degree of agreement. Thus, it appears that the reliability of the SEVTE Accuracy score is high.[25]

In order to more efficiently examine the effects of rater severity on the reliability of the SEVTE-Accuracy score, a generalizability study (G-study) was undertaken on the total SEVTE-Accuracy Score. A G-study is a means of looking at multiple sources of variance simultaneously. In this study, the two sources of variance investigated were forms and raters. The results are presented in Table 7.

---

[25] Again, it should be remembered that the consistency of the SEVTE Accuracy score is dependent on well trained raters. In an operational program, it should be possible to exceed the reliability attained in this experimental study. Operational raters will have the benefit of being able to train using the rater training materials that were developed as part of this project. In this study, the raters approached the task of rating without the benefit of having undergone a rater training program. Ratings were done on an intermittent basis at home.

Table 7
Variance Contributions of Raters and Forms
to the SEVTE-Accuracy Total Score

Source of Variance	Variance Component Estimate	Standard Error
Persons	138.665	31.95
Forms	-.285*	.10
Raters	10.120	8.37
Persons x Forms	11.971	3.94
Persons x Raters	4.110	2.39
Forms x Raters	-.180*	.09
Residual	11.225	2.39

*A negative variance estimate is an artifact of the estimation procedure. Generally these can be regarded as equivalent to zero (Brennan, 1983, p.103).

Table 7 shows that the variance due to the raters, forms, or any two-way interactions is relatively small in comparison to the variance measured among the persons. Indeed, the second highest variance component (11.971) is only 8.6% as large as the largest component and represents only 6.8% of the total variance of 176.091. Moreover, the variance due to forms and to form by rater interaction is negligible. This argues that differences in scores due to forms are minor.

The variance components estimated in a G-study can be used in a decision study (or D-study) to estimate the reliability (generalizability coefficient) of a test under various conditions of the facets being studied. Table 8 presents the estimated generalizability coefficients given both raters and forms as sources of errors under various groupings of two forms and two raters.

Table 8
Estimated Generalizability Coefficients for the SEVTE-Accuracy Score using Different Groupings of Forms and Raters

Number of Forms	Number of Raters	Generalizability Coefficient
1	1	.84
1	2	.88
2	1	.90
2	2	.93

The results in Table 8 show that the reliability for the SEVTE-Accuracy score, when one form and one rater is used, is .84, given measurement errors due to both raters and forms. This is very high for a rater-scored test. It may be noted that the reliability using two forms and two raters (as was the case in the validation study for the development of the SEVTE) was a very high .93.

6.3. Production Section: Descriptive Statistics and Reliability of the Expression Score

Table 9 below shows the SEVTE-Expression descriptive statistics (raw scores) for the Production section of the test by form and by rater. In the Production section, only the Paragraph Translations are rated for Expression. They are rated for the three criteria that figure into the total score for Expression. These criteria are Grammar (morphology), Expression (syntax and vocabulary), and Mechanics (spelling and punctuation).

Table 9
Descriptive Statistics for SEVTE Expression: Paragraphs Subsection Form 1 (N=45) and Form 2 (N=44)

Measure	Mean	Std. Dev.	Minimum	Maximum
Grammar				
R1 F1	9.5	4.2	0	15
R2 F1	9.0	3.4	2	15
R1 F2	11.0	3.4	0	15
R2 F2	9.3	3.6	0	15
Expression				
R1 F1	6.3	3.3	1	15
R2 F1	7.5	2.8	1	13
R1 F2	8.3	2.9	0	15
R2 F2	7.1	2.9	0	13
Mechanics				
R1 F1	10.4	3.8	3	15
R2 F1	10.7	3.8	2	15
R1 F2	11.9	3.4	0	15
R2 F2	10.0	3.6	0	15
Total (for Expression production section)				
R1 F1	30.3	8.7	4	45
R2 F1	31.4	6.6	11	42
R1 F2	34.3	4.5	25.5	45
R2 F2	29.0	6.6	16.5	42

Legend: R=rater, F=form. Thus R1 F1 is the score assigned by rater 1 on form 1.

Close examination of Table 5 shows that the difficulty of the two forms is very similar. Averaging the scores assigned by both raters, we see that the Paragraph Translation Expression scores seem to be slightly lower on Form 1 for all three scoring criteria. For Form 1 grammar the mean is 9.25 versus 10.15 for Form 2. For Form 1 expression it is 6.9 versus 7.7 for Form 2. For Form 1 mechanics it is 10.55 versus 10.95 for Form 2. For the total from this section, the mean on Form 1 is 30.85; for

Form 2 it is 31.65. The total means differ by less than 1 point indicating that the Production sections of the two forms are nearly equal in difficulty as a measure of the construct of Expression.

As in the discussion of the reliability of the Accuracy scores, we will first look at interrater reliability and parallel form reliability separately. Table 10 shows the interrater reliability estimates (Pearson Product Moment Correlations) of the SEVTE Production subsections and the total Production section score for Expression. These scores are all based on the Paragraph Translation subsection of the Production section of the test. The reliability for Form 1 is listed first, followed by the reliability for Form 2.

Table 10
Interrater Reliability of
SEVTE Production Subscores and Production Total (Forms 1+2)

	Form 1	Form 2
Paragraphs-Grammar	.53	.67
Paragraphs-Expression	.81	.83
Paragraphs-Mechanics	.66	.87
Total Expression*	.83	.86

*Total for Expression is for the total of the three Expression subscores on Paragraphs only.

The interrater reliabilities for the three Expression criteria are not as high as they were for the Accuracy scores, and the interrater reliability was lower for Form 1 than for Form

2.[26] Still, the interrater reliability for the total Expression score earned on the Production section is quite respectable.

Table 11 presents the coefficient of equivalence of the total Expression scores on the Production section across forms and raters. This data is an indication of the parallel form reliability of the SEVTE across different raters.

Table 11
Coefficient of Equivalence for SEVTE Expression Scores
(Production Section only, N=43)

	Form 2 Rater 1	Form 2 Rater 2
Form 1 Rater 1	.61	.67
Form 1 Rater 2	.69	.79

This data, unlike that for the Accuracy scores, indicates that raters were less consistent in their awarding Expression scores across the different forms.

In order to examine the combined effects of rater and form interaction on the reliability of the SEVTE-Expression Production Subsection, a generalizability study (G-study) was undertaken on

---

[26] It should be noted that interrater reliability is a rater characteristic, not a test characteristic. Nevertheless, a test developer must present information on interrater reliability. In the future, the interrater reliability of the SEVTE will depend on the reliability of the individuals who score the SEVTE. Raters in the SEVTE operational program, however, will have the advantage of having available training materials that were generated as a by-product of this study. Thus, these SEVTE operational raters should exceed the reliability of raters in this developmental study. In this study, the raters approached the task without the benefit of having undergone a rater training program. Thus, the raters may have used different scoring standards at different points during the three months that they were rating the production section. Ratings were done on an intermittent basis at home.

the total SEVTE-Expression Production Score. As in the previous study, the two sources of variance investigated were forms and raters. The results are presented in Table 12.

Table 12
Variance Contributions of Raters and Forms
to the SEVTE-Expression Production Total Score

Source of Variance	Variance Component Estimate	Standard Error
Persons	29.170	7.52
Forms	-5.379*	4.41
Raters	-3.321*	4.72
Persons x Forms	6.737	2.69
Persons x Raters	-.670*	1.38
Forms x Raters	10.563	8.81
Residual	9.767	2.08

***The negative variance estimate is an artifact of the estimation procedure. Generally these can be regarded as equivalent to zero (Brennan, 1983, p.103).**

Table 12 shows that the variance due to the raters, forms, and person by rater interaction is relatively small in comparison to the variance measured among the persons. However, there are some large variances due to interactions. The forms by rater interaction, the second highest variance component (10.563), is 36% as large as the largest component and represents 19% of the total variance of 56.237. This indicates that raters were not consistent in the way they awarded points across the two forms, as the data in Table 11 also suggests. This can be illustrated by comparing the total Expression Production means in Table 9. On Form 1, Rater 2 is more lenient (31.4 versus 30.3 for Rater 1). On Form 2, however, Rater 1 is more lenient (34.3 versus

29.0 for Rater 2). In addition, the variance component due to person by form interaction is also noteworthy. This indicates that to some extent examinees were not performing consistently across the two forms. Finally, the residual amount of variance, which includes the three-way interaction of persons by forms by raters and any random variance, is also relatively large. These results indicate that further training of raters on rating the paragraphs for Expression scores will be necessary in the operational program of the SEVTE and that the reliability for Expression score may be low.

Table 13 presents the estimated generalizability coefficients from a D-study produced by the variance components estimated above given both raters and forms as sources of errors under various groupings of two forms and two raters.

Table 13
Estimated Generalizability Coefficients for the SEVTE-Expression Production Score using Different Groupings of Forms and Raters

Number of Forms	Number of Raters	Generalizability Coefficient
1	1	.64
1	2	.71
2	1	.78
2	2	.83

The results in Table 13 show that the reliability for the total SEVTE-Expression score on the Production section, when one form and two raters are used, is .71, given errors due to both forms and raters. Although this is only moderate, two things should be noted. First, this score makes up only part of the SEVTE total Expression score since the multiple choice section is also included in it. Second, the reliability using two forms and two raters (as was the case in the validation study for the development of the SEVTE) was an acceptable .83.

The final total SEVTE Expression score is a composite of an examinee's score on the Multiple Choice section of the test and the Production section total, discussed above. Most of the points that can be earned by an examinee in the SEVTE Expression score are earned in the Multiple Choice section; i.e., the Expression score is the sum of the three subscores in the Production section (maximum of 45 points) and the MC section raw score (maximum of 60 points), as explained in section 1.3 of this report. Because the total Expression score is a composite of the Multiple Choice section score and the Production score, it is not possible to calculate a single empirical estimate of the reliability of this composite score in the same convenient way that one might do for a multiple choice test. There are, however, a number of ways of looking at the reliability of this composite score.

First, in order to examine the effects of different raters on the consistency of the composite SEVTE Expression score, we can calculate the degree of agreement in composite Expression scores when different raters score the Production section. The correlation between the composite Expression scores, when the points awarded by each rater are added to scores obtained on the corresponding MC section, is .95 for Form 1 and .89 for Form 2 (with scores for Form 2 weighted as described in section 5.2). These correlations are quite high, suggesting that the composite Expression score is quite stable across raters.

A second way is to look at the consistency of scores earned on the two different forms. This comparison produces an index known as the coefficient of equivalence or parallel form reliability. This coefficient of equivalence is represented in Table 14 below.

Table 14
Coefficient of Equivalence for SEVTE Expression Composite Scores
(N=43)

	Form 2 Rater 1	Form 2 Rater 2
Form 1 Rater 1	.78	.78
Form 1 Rater 2	.82	.83

This table depicts the four indexes of equivalence that can be calculated when each of two test forms is scored by two raters. As can be seen, the average coefficient of equivalence is about .81.

A final way to examine the reliability of the composite Expression score is use coefficient alpha to examine the reliability of the composite score formed by adding together the two part scores (MC and Production). In other words, under this procedure the two part scores are viewed as two subtests. It is appropriate to do this when the subtests of a composite are parallel. When subtest of a composite are parallel, then coefficient alpha can be referred to as the coefficient of precision (Crocker and Algina, 1986, p. 121), which is an estimate of test-retest reliability. An example of parallel subtests would be an essay test score that is a composite score based on two ratings. When the subtests or part scores are not parallel, coefficient alpha must be thought of as a lower bound estimate of this coefficient of precision.

In applying coefficient alpha to the SEVTE Expression scores, it is appropriate to average the production section scores awarded by the two raters used in this study. This mean score on the production section gives us the best estimate of the scores that would be awarded by any other rater who may score this test. Calculated in this manner, coefficient alpha is .76 for Form 1 and .53 for Form 2, with unweighted scores being used for Form 2. Since the MC section and the Production section are so different, they cannot be considered parallel subtests. Thus, it is not surprising to find lower bound estimates of the coefficient of precision for the SEVTE in this moderate range.

7. Examining the Validity of the SEVTE

According to the Standards for Educational and Psychological Testing (American Educational Research Association, et al., 1985), test validity refers to "the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores" (p. 9). Validity is demonstrated by an accumulation of evidence that supports the claim of validity for a particular test. Some of this evidence is empirical. Other evidence may be qualitative, in that it deals with the content of the test, or it may be theoretical, in that it deals with a theory about the nature of the trait being measured by the test. In the case of the SEVTE, the central validity concern is the claim that the test is a measure of the ability to translate a written text in Spanish into correct and appropriate English.

Traditionally, three types of validity are usually identified according to how the evidence was gathered. These are content validity, criterion-related validity, and construct validity. Construct validity, which "focuses primarily on the test score as a measure of the psychological characteristic of interest" (AERA, et al., 1985, p. 9), may be understood to subsume the other two types; i.e., content and criterion-related validity are also evidences of the construct validity of a test. Thus, construct validity is of central interest. We will work toward a discussion of the construct validity of the SEVTE, by beginning with an analysis of its content validity. Subesequently, we will examine the construct validity of the test more directly, through analyses of the trait that is being measured by the test. Finally, we will examine the criterion-related validity of the SEVTE by considering its relationship to success at translating and to other measures of language proficiency.

7.1. Content Validity

Content validity is evidence that demonstrates the degree to which the sample of items, tasks or questions on a test are representative of the domain of content that could be tested. In the case of the SEVTE, evidence for its content validity is found in the tasks examinees are asked to perform to demonstrate their ability to translate from Spanish to English.

First, the Multiple Choice section involves two general tasks required of Spanish/English translators: recognizing whether a proposition in Spanish is rendered into English with appropriate expression, and recognizing errors in written English. Clearly, the ability to select the appropriate word or phrase from among the many that could be available or correct in other contexts is a skill that a translator must have. A translator uses this ability to recognize infelicities in his or her work in order to revise it successfully. In addition, the ability to recognize errors in English is important because the translator must be able to revise his or her first draft so that it represents appropriate English expression. Otherwise, the translator's English rendition can be accurate in terms of the rendition of the content of the source document, but it will still appear to be a translation.

The SEVTE tests these two abilities through 60 Multiple Choice items: 35 Words or Phrases in Context (WPC) items and 25 Error Detection (ED) items. WPC items test a wide variety of points of English and Spanish grammar. These points include subject-verb agreement, verb tenses, pronouns, prepositions, gender, and word order. They also test a range of Spanish-English vocabulary, including nouns, verbs, adverbial and adjectival phrases, and false cognates. Each item on each of the two forms of the test focuses on the same or nearly the same aspect of grammar or vocabulary. The 25 ED items include errors of grammar, word order, vocabulary, punctuation or spelling in English only. Thus, of the seven criteria included in the Translation skill level descriptions (accuracy, grammar, vocabulary, style, tone, spelling, and punctuation) developed for this project, these Multiple Choice items test all except style and tone.[27] (For additional information relevant to the content validity of the Multiple Choice section, see the content analysis in Appendix D.)

Second, apart from the ability to identify correct and incorrect expression, the ability to produce a correct translation is clearly required of a translator. The ability to produce a correct translation is assessed through 28 direct production tasks. 15 of these tasks involve the translation of a word or a phrase within a sentence, called Word and Phrase Translation (WPT); 10 involve the English translation of complete Spanish sentences (called Sentence Translation or ST) that range in length from 12 to 25 words; and 3 tasks require Paragraph Translation (PT), the ability to produce an English translation of a paragraph written in Spanish. The three paragraphs range in length from approximately 80 to 120 words.

---

[27] One way that vocabulary is tested is through the mistranslation of words. Mistranslation involves both the vocabulary and accuracy aspects of the SLDs. Thus, the construct of Accuracy is partly represented in the content of the multiple-choice section.

The 15 Word and Phrase Translation (WPT) items and the 10 Sentence Translation (ST) items present examinees with a variety of problems in vocabulary, idioms, grammar (morphology) and syntax. We judged the sentences to range in difficulty from 2+ to 4+ on the Translation Skill Level Descriptions, based on the frequency and complexity of language they employ and the difficulty that language presents to the translator.[28] The items in each section are grouped by order of the perceived difficulty of the sentence on the FBI\CAL SLDs. Corresponding items on each of the two forms are parallel in content and perceived difficulty.

For WPT items, item developers relied on their expertise as translators and as language teachers in order to develop appropriate items. They created items that test aspects of the language that present special difficulty when translated to the target language, often cases where there is no direct equivalent. For example, the expression "priced in the teens," has no direct equivalent, and use of the dictionary would not be helpful. In this case, the translator must use his knowledge of both languages to construct an appropriate translation.

---

[28] As indicated by Stansfield and Liskin-Gasparro in Duran et al. (1985), it is heretical to classify decontextualized language, such as words, phrases, or sentences on the ILR scale. Still, for research or training purposes it is sometimes necessary to do this. An appropriate disclaimer of these difficulty levels is noted here.

The ST items were constructed to include grammar problems that have traditionally created difficulties for translators and language students because of a lack of congruence between the two languages. Such problems include pronouns, verb tenses and sequences of verb tenses, use of negatives, possessives, prepositions, and nontemporal verb forms, such as infinitive, gerund, and past participle.

The first Paragraph Translation (PT) text is a newspaper account, using mature vocabulary and syntax, of a crime that occurred in a Spanish-speaking country. The subject of the crime is airplane hijacking or sabotage, depending on the form of the test. This text was judged to be a low level 3 text based on the ILR skill level descriptions for reading.

The second PT text is political/philosophical in nature. It deals with either the Armed Forces or ecology. The scoring guidelines (see Appendix F) are based on the Translation Skill Level Descriptions developed for this project. The difficulty level of this text was judged to be at 3+.

The third PT text is a law or a legal interpretation of a law. The scoring guidelines (Appendix F) are based on the Translation Skill Level Descriptions developed for this project. The guidelines for scoring all the paragraphs include nearly all of the criteria included in the Translation Skill Level Descriptions. The difficulty of this document is considered to be at the 4+ or 5 level on the ILR skill level descriptions for reading. Thus, the third text is clearly the most difficult.

The entire Production section is scored using scoring guidelines that are based on the level descriptions in the FBI/CAL Translation Skill Level Descriptions (see section 4.2 and Appendix I). These descriptions were developed over a period of six months and represent a consensus of the experience of experienced translators and translation test evaluators.

The text material that appears on the SEVTE was influenced by the results of the survey of FBI translation needs (see Appendix Q and section 1.3 of this report). This questionnaire was responded to by 28 Language Specialists and agents. The results indicated that the written materials the respondents most often deal with involve politics, narcotics, terrorism, foreign counterintelligence, written laws, theft, and organized crime. Many of the SEVTE texts were actually provided by the FBI, and those found by CAL staff were judged relevant by FBI Language Specialists. Texts found by CAL staff were taken from two sources: public documents such as newspapers and magazines, and documents that item writers actually have translated in their own translation work. The texts taken from public documents were guided by sample texts provided by the FBI, especially in terms of vocabulary. These texts, as well as the texts that item writers had previously translated on the job, were edited slightly to make them more suitable for these tests. The third paragraph, which is a legal document written in appropriate jargon, (sometimes referred to as "legalese" among government Language Specialists) was supplied by the FBI for both forms of the ESVTE. In order the make the SEVTE as parallel as possible to the ESVTE, CAL staff located similar legal documents in Spanish for the SEVTE.

It is interesting to examine the responses of the validation study subjects (Special Agents, Contract Linguists, Language Specialists and others) to the exam feedback questionnaire they completed after taking the test (see Appendix L). On this questionnaire, 50% either agreed or strongly agreed with the statement, "The material in the exams was representative of the types of written documents I might encounter in my work." Another 50% either disagreed or disagreed strongly with the statement. It is difficult to interpret this data in terms of job relevance. Judgments of the job relevance of a test are highly dependent on the relationship between the test and the job of the individual subject, and the subjects in the sample varied greatly in the agency they worked for and in the job they performed. It must be remembered that within the sample of 58 examinees, 22% were beginning and intermediate level CIA Spanish language students who would not have ever translated such material, 26% were FBI Special Agents, 19% were FBI Language Specialists (or Contract Linguists who do similar work), 17% were FBI support staff, and 10% were members of the Houston Police Department. The SEVTE was designed with the knowledge that it would be taken principally by potential and current Language Specialists and others who might wish to demonstrate the ability to do the type of translation that Language Specialists regularly do. Yet Language Specialists made up only 19% of the validation study sample. Under the circumstances, the responses to the job relevance question on the exam feedback questionnaire are not as negative as might have been expected.

One of the subjects wrote on the questionnaire: "The vocabulary and material given in this test do not represent the material we are required to work with in the field. This is geared mainly to the FCI LS's (foreign counterintelligence work and Language Specialists)--not those of us working in the criminal/drug cases." This telltale comment, apparently written by a Special Agent, represents the perception that the test reflects written material that FBI Special Agents are not normally asked to translate. Most written translation is done by Language Specialists. Thus, although critical of the test, the above comment reflects the perception that the test is relevant to the work of an FBI Language Specialist.

At the same time, it is noteworthy that there was a more general agreement that the test measured translation ability. 59% percent of the subjects either agreed or strongly agreed with the statement "There was sufficient opportunity for me to demonstrate my ability to translate from Spanish to English." It may be that the 41% who disagreed with this statement did so because they felt unduly restricted by the time constraints of the testing situation; over half (53%) of the subjects felt the length of time given for the production section was "too short," and none felt it was "too long." 47% felt it was "about right." (It may be noted that on the Multiple Choice section, examinees were markedly more positive about the length of time given, with 92% indicating it was "about right," and 8% responding that it was "too short.")

In interpreting the responses to the examinee questionnaire, it is important to note that approximately 15% of those who took the SEVTE in the validation study had received scores of 2+ or less on the Spanish OPI (see section 4.4.3 above). These subjects may have understandably felt pressured by the exam time constraints, since nearly all of the tasks on the test were above their level of ability. On the other hand, those subjects whose proficiency was very high may not have had sufficient time to revise their translations. Indeed, several of the examinees indicated this to test administrators, who in turn reported it to CAL on the test administrator report form. Because of this, CAL has recommended that the amount of time allowed for completing the Paragraph Translation subsection be increased from 37 to 48 minutes; i.e., 11 minutes more than examinees in the validation study sample were permitted. This may have the effect of raising scores on the test somewhat.[29]

In general, the implications for test validity of the responses to the examinee questionnaire are lessened by the fact that a) most examinees in the validation sample were not Language Specialists, b) because of this, many had low ability in written translation, and c) the test was too speeded. This last problem has been corrected on the current form of the test by increasing the time limit for the Paragraph Translations from 37 to 48 minutes.

7.2. Construct Validity

Traditionally, validity has been defined as the degree that a test measures what it claims to measure. Evidence of validity has been divided into three types: content validity, construct validity, and criterion-related validity. However, during the past 15 years, validity has come to refer to the inferences that can legitimately be made from test scores for a particular type of examinee and for a particular purpose. Similarly, construct validity has become synonymous with validity itself (Messick, 1980). Because of this, the same definition is also the contemporary definition of construct validity. However, within the context of the validity section of this report, we have made use of the traditional division of kinds of validity in order to organize a fairly complex presentation of the evidence for validity that was gathered. Thus, we will now consider the more limited, traditional definition of construct validity; that is, the dimensions of ability that are being measured by the test.

---

[29] The general increase in the test scores that may be obtained by increasing the time available to examinees to complete the test should be viewed positively. It is likely that if scores do increase under extended time limits, this will be due to a reduction in test speededness, and the scores will be more accurate. For additional information, see Appendix P.

In the introduction to this report we identified and described two dimensions of translation ability: Accuracy and Expression. We discussed how these dimensions evolved from our efforts to develop Translation SLDs, from our research on the Listening Summary Translation Exam, and from our initial scoring of the SEVTE test papers. These two dimensions of translation ability were strongly supported by the results of our analyses of the SEVTE test data. Thus, we begin this analysis of the construct validity of the SEVTE by stating that the test claims to measure overall translation ability, but that it divides this ability into two dimensions (Accuracy and Expression) and it claims to measure each. Accuracy is the degree to which the information in the source document is conveyed in the target document. Errors in Accuracy include the misrepresentation or deletion of information in the source document, or the inclusion of information that was not in the source document. Expression, on the other hand, focuses on the appropriateness of the language used in the target document.

When a test measures two distinct dimensions, the measures of those should demonstrate some unique score variance. Thus, while the measures may be related, they should be distinguishable. Table 15 below presents the correlations between the total scores for Accuracy and Expression for Forms 1 and 2 of the ESVTE.

Table 15
Correlations between Mean Total Expression and Accuracy Scores on Form 1 and Form 2
(n = 44)

	TOTEXPF1	TOTEXPF2	TOTACCF1	TOTACCF2
TOTEXPF1	1.00			
TOTEXPF2	.83	1.00		
TOTACCF1	.74	.63	1.00	
TOTACCF2	.75	.73	.90	1.00

Legend:
TOTEXPF1 = Total Expression Score, Form 1
TOTEXPF2 = Total Expression Score, Form 2
TOTACCF1 = Total Accuracy Score, Form 1
TOTACCF2 = Total Accuracy Score, Form 2

As can be seen in table 15, the correlation between these two total scores for Form 1 is .74, while for Form 2 it is .73. These moderate correlations suggest that the two subscores are measuring different but related abilities. This finding is further corroborated by examining the correlation between the two scores that claim to represent the Accuracy dimension and the two scores that claim to measure the Expression dimension. Note that the correlation between the Accuracy score on Form 1 and the Accuracy score on Form 2 is .90. Similarly, the correlation between the Expression total score on Form 1 and the Expression total score on Form 2 is also .83. These correlations between measures of the same dimension clearly exceed the correlations between the measures of different dimensions mentioned above. Thus, since each measure correlates more highly with a measure of the same dimension than it does with a measure of a different dimension, it is clear that the SEVTE measures two dimensions of translation ability. Correlations of this nature suggest that one score cannot serve as a substitute for the other. Because individual examinees often have different ability levels on each, both Accuracy and Expression need to be assessed on a Spanish to English translation test for this population. However, because the two measures show moderately high intercorrelations, each subscore is also a measure of the global trait being measured by the test.

We will now turn to a discussion of criterion-related validity. This discussion provides a better understanding of the global trait being measured and how it relates to other relevant traits.

7.3. Criterion-related Validity

Criterion-related validity is evidence that "demonstrates that test scores are systematically related to one or more outcome criteria" (AERA, p. 11). For example, if supervisors ratings of employees' translation ability were available, then it would be important to see how scores on the ESVTE and supervisors ratings compared. Unfortunately, the Special Agent in Charge at each local FBI office is rarely able to rate the translation ability of Language Specialists or Special Agents, because a variety of languages may be represented in each field office.

Thus, an appropriate existing criterion variable was not available to the authors of this study.

In an effort to remedy this situation, we constructed two concurrent measures that can serve as a variable for determining criterion-related validity. The concurrent criterion-related variables are described below.

Concurrent Criterion-Related Measures

Overall FBI/CAL Expression and Accuracy Scores (EXPFBICAL and ACCFBICAL). After the two raters in the validation study assigned analytical scores to each section of the production section of the SEVTE, they assigned each examinee two overall scores on the FBI/CAL Translation SLDs: one for Expression and one for Accuracy, based on the examinee's performance on the Sentences and Paragraph subsections of the Production Section. Each examinee took two forms. Thus, each examinee's overall FBI/CAL Expression and Accuracy score is the average of four ratings (two raters by two different forms). These overall FBI/CAL Expression and Accuracy scores were obtained for all subjects. They provide two measures of criterion-related validity.

The data on the two concurrent criterion-related validity measures provide a basis for assessing the criterion-related validity of the SEVTE. Correlations between the Total Accuracy and Expression scores on each form of the SEVTE with these concurrent measures are presented in Table 16 below.

Table 16
Correlations of the SEVTE Scores with Overall Rating of Translation Ability (Numbers of Paired Scores in Parentheses)

	EXPFBICAL	ACCFBICAL
EXP1	.88* (44)	.76* (44)
EXP2	.89* (43)	.75* (43)
ACC1	.78* (44)	.89* (44)
ACC2	.83* (43)	.92* (43)

* $p < .0001$

Before beginning a discussion of these relationships, it is appropriate to consider the validity and reliability of the two measures of criterion-related validity (EXPFBICAL and ACCFBICAL).

As indicated in the description of the FBI/CAL overall Expression and Accuracy ratings, after scoring each paper analytically, the raters then referred to the FBI/CAL Translation SLDs to determine an appropriate holistic rating for each examinee based on his or her performance on the Sentences and Paragraphs subsections of the Production section of the test. This holistic rating is a rating of overall translation ability based on performance in translating 10 challenging sentences and three paragraphs varying in difficulty. Thus, this holistic rating can be considered a performance-based assessment of translation ability. Its validity as such is limited slightly by the fact that of the four ratings that go into this composite holistic rating (two ratings on each form), two were awarded by the same rater that scored the form correlated in Table 16 with the holistic rating. Thus, two of the ratings are not wholly independent. However, the other two ratings were based on success at translating different texts. In this case, the different texts were the sentences and paragraphs appearing on the other SEVTE form. While one approach might have been to use the FBI/CAL skill level assigned by the two raters who scored the other form as the criterion variable (as discussed in footnote 30), we chose to combine all four ratings from the two forms into a single indicator of translation skill level in this study. This composite rating has the advantage of being based on twice as many performance tasks, (20 sentences and six paragraphs) and twice as many ratings of translation skill level; that is, four ratings instead of two ratings. Thus, this composite rating of translation skill level can be considered to be both more reliable and more valid because of the number of tasks and evaluations (ratings) on which it was based.

In order to determine the reliability of the criterion variables, i.e., the composite FBI\CAL overall rating of translation ability for Accuracy and Expression, a Generalizeability (G) study was performed on the data that went into the composite rating. A G study is a statistical technique in which the contributions of various factors (facets) to the total variance of the test scores are estimated. For this particular study, we wanted to estimate how much of error variance was contributed by the raters and the forms. (The forms are the two different samples of translation ability that are elicited by SEVTE Form 1 and Form 2.) There were 44 examinees and two raters involved in the G study. Thus, both criterion variables (EXPFBICAL and ACCFBICAL) received four ratings. In our study, we wanted to estimate the generalizability coefficient for the average translation ability rating for Expression and Accuracy when two ratings on two forms were used to construct the average. The G coefficient is an estimate of reliability, based on the ratio of the variance of the objects of measurement (in this case persons) over that variance plus error variance due to forms, raters, and their interactions. The results of the studies indicated that the G coefficient for the EXPFBICAL rating is .85 and the G coefficient for the ACCFBICAL rating is .88. These G coefficients may be considered the reliability of these two criterion variables.

Returning now to Table 16, the correlations between the criterion variables (EXPFBICAL and ACCFBIILR) and the SEVTE Expression and Accuracy scores are consistently high. Of the eight correlations shown, the lowest is .75 and the highest is .92. The fact that scores on the SEVTE correlate highly with overall translation skill level ratings strongly supports the validity of the two scores.

Further analysis shows that the correlations improve as one might expect. The correlation between the SEVTE Expression score with the Expression criterion variable (EXPFBICAL) is .88 for Form 1 and .89 for Form 2. This is strong evidence of the validity of the SEVTE Expression score.

Similarly, the correlation between the SEVTE Accuracy score and the Accuracy criterion variable (ACCFBICAL) is high also: .89 for Form 1 and .92 for Form 2. This is strong evidence for the validity of the SEVTE Accuracy score.[10]

7.3. Convergent/Discriminant Construct Validity

---

[10] Although we chose to use the average of the four overall FBI/CAL translation ability level ratings here as a criterion variable, it is interesting to consider the correlations between the SEVTE Expression and Accuracy scores on one form and the overall FBI/CAL translation ability level ratings assigned by the raters based on the examinee's performance on the other form. In this case, the other form is a totally independent criterion variable. That is, the rating is based on the examinee's performance on other translation tasks like the ones that an examinee would have to perform on the job.

Here the validity coefficients are also quite good. The correlation between the SEVTE Expression total based on Form 1 and the average of the two overall FBI/CAL translation skill level ratings assigned based on Form 2 Sentences and Paragraphs is .83. Similarly, the correlation between the Expression total based on Form 2 and the average of the two overall FBI/CAL translation skill level ratings assigned based on Form 1 Sentences and Paragraphs is .81.

The correlation between the SEVTE Accuracy total based on Form 1 and the average of the two overall FBI/CAL translation skill level ratings assigned based on Form 2 Sentences and Paragraphs is .83. Similarly, the correlation between the Accuracy total based on Form 2 and the average of the two overall FBI/CAL translation skill level ratings assigned based on Form 1 Sentences and Paragraphs is .81.

Again, it must be remembered that these overall FBI/CAL translation skill level ratings are less reliable than those included in table 16. The G study showed the G coefficient with one form and two ratings to be .77 for EXPFBICAL and .79 for ACCFBICAL.

Because the evidence in Table 16 so clearly supports the validity of the SEVTE as a measure of Spanish-English translation ability, a fuller discussion of evidence for the construct validity of the test is warrented. Such a discussion can be obtained by considering the convergent/discriminant nature of the correlations between the SEVTE and other measures theoretically related to the construct of interest. In such a discussion, an expected correlation of the test with each variable is analyzed and discussed. Some criteria will be expected to show a strong relationship with the test whose validity is being examined, while other criteria will be expected to show a weak correlation, or to not correlate at all, or even to correlate negatively. We will make use of the convergent/discriminant validity approach here in order to fully examine the construct validity of the SEVTE.

In an effort to attain further understanding of the construct measured by the SEVTE, two concurrent measures were collected. These concurrent measures are described below.

1. A self-rating (SPENSELF and ENSPSELF). CAL developed two questionnaires that asked subjects a) with what types of documents they had experience translating from Spanish into English and English into Spanish; and b) if they had experience, to rate their translation ability of these documents as either "Limited," "Functional," "Competent," or "Superior." These questionnaires were administered to the subjects immediately preceding the administration of the first part of the corresponding test. A copy of these questionnaires is contained in Appendix N. Each subject's responses to these two questionnaires were converted into self-rating scores (Spanish into English = SPENSELF; English into Spanish = ENSPSELF) by first awarding points to each item that subject rated (1 for "Limited," 2 for "Functional," 3 for "Competent," 4 for "Superior," with N/A

receiving no value) and then calculating the mean response to all items for which he or she provided a self-rating.

In addition, data were collected, where available, on six nonconcurrent tests that had been administered within one to eight years of the study.

Previously Administered Measures

1. A Spanish OPI score (SPANSPK). An oral proficiency interview (OPI) score for Spanish was collected for as many subjects as possible. Although this is not a wholly adequate criterion variable, it is relevant to translation ability. Speaking proficiency assumes and is moderately correlated with Spanish reading proficiency. Correlations between the two skills typically are between .50 and .75. Thus, on a theoretical basis, it was decided that the OPI score could be used to provide additional evidence of criterion-related validity. For all ILR scores in this study, the following conversion was used for purposes of empirical analyses:

ILR Score	Numerical Score
0+	0.8
1	1.0
1+	1.8
2	2.0
2+	2.8
3	3.0
3+	3.8
4	4.0
4+	4.8
5	5.0

2. Other test scores. Other scores that measure possibly related constructs were collected as possible. None of these scores could be collected for all the subjects, however. These scores, the number of subjects for which they were collected, and their descriptive statistics are given below, together with the same information on all of the measures.

```
Measure   N        Mean      Std. Dev. MinimumMaximum
-------   -        ----      --------- --------------
EXPFBICAL 44       2.86      0.67      1.304.65
ACCFBICAL 44       2.58      0.72      0.904.25
SPENSELF  43       2.89      0.67      1.04.0
ENSPSELF  35       2.90      0.62      1.04.0
SPANSPK   36       4.14      0.98      2.05.0
DLPTLIST  28       52.75 5.06          3960
DLPTREAD  28       53.25 6.54          3060
ENGSPK    17       4.21  0.60          3.05.0
SPENTRAN  17       3.45      0.96      2.04.8
ENSPTRAN  17       3.29      0.65      1.84.0
----------------------------------------------------------
Key
---
EXPFBICAL Overall ILR expression score.
ACCFBICAL Overall ILR accuracy score.
SPENSELF  Average score on the Spanish into English Verbatim
          Translation Ability Self Assessment Questionnaire.
ENSP[illegible]  Average score on the English into Spanish Verbatim
          Translation Ability Self Assessment Questionnaire.
SPANSPK   An OPI score for Spanish.
DLPTLIST  The listening section of the Defense Language Institute
          Placement Test.  Maximum possible score = 60.
DLPTREAD  The reading section of the Defense Language Institute
          Proficiency Test.  Maximum possible score = 60.
ENGSPK    An OPI score for English.
SPENTRAN  An ILR score on the current FBI Spanish into English
          verbatim translation exam.
ENSPTRAN  An ILR score on the current FBI English into Spanish
          verbatim translation exam.
```

Relationships between scores on these measures and scores on the SEVTE were calculated in order the examine the convergent/discriminant validity of the SEVTE.

7.4.1. Convergent Validity

Correlations between the Total Accuracy and Expression scores on each form of the SEVTE with the concurrent measures are presented in Table 17 below. (Note that the SEVTE scores in this table represents a composite of the two ratings. In addition, examinees were not penalized if they did not attempt a paragraph due to lack of time.) The number of subjects involved in the correlation is also given, since not every subject had a score on every measure; i.e., the numbers in parentheses represent the number of subjects who had a score on both measures being correlated. The magnitude of the Ns should be considered in making interpretations. Larger Ns allow a greater degree of confidence in the indicated relationship. In general, none of the Ns are large, suggesting that the correlations should not be considered stable.

Table 17
Correlations of the SEVTE Scores
with Other Available Measures
(Numbers of Paired Scores in Parentheses)

	SPENSELF	ENSPSELF	SPANSPK	DLPTLIST	DLPTREAD	ENGSPK	SPENTRAN	ENSPTRAN
EXP1	.43* (43)	.38* (35)	.04 (36)	.56* (28)	.65* (28)	.50* (17)	.50* (17)	.49* (17)
EXP2	.28 (42)	.25 (34)	-.07 (35)	.43* (27)	.30 (27)	.51* (17)	.50* (17)	.50* (17)
ACC1	.63* (43)	.42* (35)	.47* (36)	.76* (28)	.70* (28)	.47 (17)	.57* (17)	.75* (17)
ACC2	.59* (42)	.53* (34)	.36* (35)	.62* (27)	.60* (27)	.53* (17)	.48 (17)	.68* (17)

* $p < .05$

We will now discuss the relationships in the Table 17, refering again, when appropriate to the data in Table 16. The accuracy of this discussion is tempered by the fact that no reliability statistics are available on any of these criterion measures. Even though this is the case, since this is the only data available, there is no other option than to examine and interpret the suggested relationships. Since the magnitude of these relationships is attenuated to the extent that the tests are less than perfectly reliable, one can generally assume that the relationships are at least as strong as are indicated here. On the other hand, the reliability of the SEVTE scores does not pose a problem, since all the SEVTE reliabilities are quite high. (See sections 6.2 and 6.3.)

First, it is most notable that there were moderate correlations, most of them significant, between the SEVTE Total Accuracy score and all the criterion variables. The correlations between the SEVTE Expression score and the criterion variables were usually not as high as the correlations for the Accuracy score, and they are not always significant. This supports the centrality of the Accuracy score in the measurement of translation ability.

Accuracy is the degree to which the information in the source document is conveyed in the target document. Errors in Accuracy include the misrepresentation or deletion of information in the source document, or the inclusion of information that was not in the source document. Expression, on the other hand, focuses on the appropriateness of the language selected for use in the target document.

In the tables above, we would expect a positive correlation between the SEVTE Accuracy score and the Spanish into English self-assessment of this ability (SPENSELF). These correlations, depicted in the left column of Table 17 above, are

.63 for Form 1 and .59 for Form 2. These moderately strong correlations support the validity of the SEVTE Accuracy score. The lower correlations between SPENSELF and SEVTE Expression (.43 and .28), suggest that factors other than the ability to translate the information, i.e., English writing ability, may play a larger role in the Expression rating. Again, no data are available on the reliability of the SPENSELF questionnaire.[31]

---

[31] The question of the reliability of the questionnaires used to calculate each subject's self-assessment score deserves some comment here. When dealing with the internal consistency reliability of a measurement instrument, the estimated reliability coefficient is an indication of the extent to which items comprising the measure are tapping into the same underlying trait or ability. This assumes that each item was written to measure this trait or ability, and that all examinees would answer all items.

The nature of the two questionnaires from which self-assessment scores were calculated here was somewhat different in that each subject gave a self-rating only to a subset of the "items." These "items" were the document types with which he or she had experience. In the vast majority of cases, subjects did not have experience in translating all the document types; thus, self-rating scores were sometimes based on only 3 or 4 responses. The response on the other "items" was "Not Applicable," to which no reasonable numerical value could be assigned; "Not Applicable" means that the subject does not translate such document types.

When missing data occurs in a questionnaire database, there are several ways to deal with the problem under certain circumstances. Inadvertently missing data may be replaced by an estimate of that subject's response to the item, such as using his or her mean score on items answered or the mean response of all subjects answering that item. On certain measures, such as on an attitudinal questionnaire, a missing value may be appropriately interpreted as the subject's having no opinion or not caring about the issue in the item, and a missing value can then be replaced by a neutral response.

Had we been able to treat these responses as missing data, there would have been several ways to estimate the reliability of the two questionnaires. However, on the questionnaires used here, a response of "Not Applicable" is not missing data. To replace these responses with a numerical value (such as the subject's mean response) is contrary to the subject's own rating of "Not Applicable" to that "item" (document type). Furthermore, even if it were appropriate to treat the response as missing

The correlations between the SEVTE and the self-rating of ability to translate each of the 12 types of documents included on the SPENSELF questionnaire are found in Appendix N. Given the relatively small proportion of Language Specialists in the sample, it is possible that most examinees did not have much experience translating such documents on the job. An attempt was made to correct for this in the design of the questionnaire by telling people in the instructions, "If you have never translated a particular type of document, please mark N/A (not applicable)." While almost all subjects completing the questionnaire (43) indicated that they translated correspondence (letters) (98%), the mean number of documents responded to of the 12 document types was 7.79. While all document types received at least a 47% response, the average examinee responded N/A to about a third of the document types. Thus, it may be inferred that translation of documents other than letters is performed rarely by most examinees and consequently that most examinees may have not have had a valid basis for making judgments of their ability.

It is worthwhile to consider the correlations between SEVTE scores and the self-ratings of ability to translate the 12

---

data, making a large number of replacements as would be required here, would inflate reliability by increasing interitem consistency in proportion to the number of responses of "Not Applicable" that were replaced by each subject's mean response. The resultant estimate of reliability would thus be spuriously high and it would not be interpretable.

document types included on the Spanish-English Self-Assessment Questionnaire. All of the 24 correlations between the SEVTE Accuracy score for Forms 1 and 2 and the 12 document types were significant. The correlations ranged from .74 to .42. The highest correlations were with the ability to translate foreign diplomatic reports (.73 and .74),[32] depositions (.73 and .72), foreign counter-intelligence status/evaluation reports (.65 and .57), correspondence (.59 and .64), letters rogatory (.54 and .62), police reports (.56 and .56), and news editorials (.57 and .51). These correlations, individually and as a whole, provide evidence of the convergent validity of the SEVTE Accuracy score. The fact that the correlations are so similar for the two forms also bodes well for the comparability of the two forms. That is to say, they appear to measure the same construct.[33]

Another overall measure of translation ability is the FBI's current Spanish to English translation test (SPENTRAN). (See column 7 above.) The SEVTE Accuracy and Expression scores correlated moderately with this test (.48 to .57) for the 17 examinees for whom scores on this test were available. One must remember that the FBI is not satisfied with the reliability and validity of this test.[14] Thus, the lack of a high correlation with the SPENTRAN should not be a source of concern. Under the circumstances, the magnitude of this correlation is acceptable.

---

[32] The first correlation in parentheses is with the Accuracy score for Form 1 and the second is with the Accuracy score for Form 2. All of the correlations and the Ns on which they are based are available in Appendix N.

[33] The correlations between the 12 document types and the SEVTE Expression score were lower and less than half were statistically significant.

Theoretically, the ability to translate from Spanish to English should require reading ability in the language of the source document, which is Spanish. The measure of Spanish reading ability used here was the DLPT Reading subtest. The SEVTE Accuracy score showed moderately high correlations (.70 and .60) with the DLPTREAD, which indicates that it is sensitive to Spanish reading proficiency. One would expect the SEVTE Expression score to be less related to Spanish reading ability. The Expression correlations with DLPTREAD (.45 and .30) show that this was indeed the case, and in the case of Form 2, the correlation was not significantly different from zero.

Another measure of Spanish ability available was the Spanish OPI score (SPANSPK). There was a moderate correlation (.47 and .36) between SPANSPK and the SEVTE Accuracy, confirming that Spanish language ability is related to the ability to translate information from Spanish to English. However, there was no correlation (.04 and -.07) between SPANSPK and SEVTE Expression. This indicates that Spanish speaking ability is not related to the ability to translate a Spanish language text using appropriate English written expression. This is as expected, and supports the use of two separate scores for the SEVTE.

---

[14] No evidence of the reliability of this test has ever been gathered.

English proficiency should also be necessary to translate from Spanish to English. The only measure of this proficiency available was the English OPI score (ENGSPK). The correlation between English speaking proficiency and SEVTE Accuracy (.47 and .53) was about the same as it was for Spanish speaking proficiency. In addition, the ENGSPK correlation with Expression (.50 and .51) is about equal in magnitude to its correlation with Accuracy, suggesting that both SEVTE scores are related to English proficiency. It may be noted here that whereas SPANSPK was not correlated to total Expression scores, ENGSPK was. This is understandable, since English speaking ability can be expected to correlate with English writing ability, whereas Spanish speaking ability would not be expected to correlate with English writing ability.

7.4.2. Discriminant Validity

Another criterion-related approach to establishing construct validity is to consider all the measures as a whole and contrast the correlations. First, one begins with the measures that one would expect to show a low correlation with the SEVTE. Then, one contrasts these measures with the correlations for the measures that one would expect to correlate more highly with the SEVTE. If the correlation with the variables expected to be more relevant is indeed greater, then this is evidence of discriminant validity. Thus, one examines the magnitudes, the differences, and the direction of the differences of the correlations, to see if they fullfill a priori expectations. This process establishes the discriminant validity of the test under consideration. Using this approach, the data from the validation study is usually, although not always, supportive of construct validity of the SEVTE as a test of Spanish to English translation ability.

First, we will begin by comparing the SEVTE with the two concurrent criterion-related validity variables in Table 16. These variables are the composite rating of translation skill level assigned by the raters after analytically scoring the production section of the test. In Table 16, we see that SEVTE Expression score correlates more highly with the translation skill level for Expression (EXPFBICAL) than it does with the translation skill level for Accuracy (ACCFBICAL) (.88 and .89 versus .76 and .75). We also see that the SEVTE Accuracy score correlates more highly with the translation skill level for Accuracy (ACCFBICAL) than it does with the translation skill level for Expression (EXPFBICAL) (.89 and .92 versus .78 and .83).

Second, we will compare the SEVTE with other measures of Spanish-English translation ability. The self assessment questionnaires (SPENSELF and ENSPSELF) completed by examinees prior to the exam are two such measures. One would expect to find a stronger relationship between SEVTE scores and the SPENSELF than between the SEVTE scores and the ENSPSELF, since the ENSPSELF is a measure of translation in the opposite direction. Columns one and two in Table 17 indicate that this turned out as expected. All four of the SPENSELF correlations are larger than the corresponding ENSPSELF correlation.

Two other such measures are the FBI's current translation tests (SPENTRAN and ENSPTRAN). One would expect a stronger relationship between the SEVTE and the SPENTRAN, since both purport to measure the ability to translate in the same direction. Such an outcome was not found, however. In two out of four comparisons, the ENSPTRAN showed the stronger correlation and in two cases there was essentially no difference. Again, one must remember that these current FBI tests are considered to have limited validity.

Another issue is the relative importance of the two languages to the two scores. One would expect the SEVTE Expression score to be more strongly related to English proficiency than to Spanish proficiency, since, on the SEVTE, the examinee actually performs in English. The one measure of English proficiency available is ENGSPK and the three measures of Spanish proficiency available are SPANSPK, DLPTLIST, and DLPTREAD. The SEVTE Expression score shows a far greater correlation with ENGSPK (.50 and .51) than with SPANSPK (.04 and -.07), which is a measure of the corresponding skill (speaking). The direction of the difference is as one would expect. SEVTE Expression also shows a higher correlation with ENGSPK than with DLPTREAD (Spanish reading) (.45 and .30), which is also as one would expect. However, the SEVTE Expression correlation with DLPTLIST is about equal to the correlation with ENGSPK, even though one would expect it to correlate higher with ENGSPK.

There is no explanation why the correlation with DLPTLIST was so high, since translation does not involve listening. Again, one must remember that the sample size for this correlation was small (N=17), and that correlations based on small Ns can vary greatly from the true correlation.

Similarly, one would expect the SEVTE Accuracy score to be more strongly related to proficiency in Spanish than is Expression.[35] The data for the three measures of Spanish (SPANSPK, DLPTLIST, DLPTREAD) show this to be the case. In fact, the difference in the correlations for Accuracy and Expression is far greater on these measures of Spanish than for three other measures in Table 17, namely ENGSPK, SPENTRAN, and ENSPTRAN.[36]

Similarly, since Accuracy, theoretically involves both languages about equally, one would expect fairly similar correlations between Accuracy on corresponding measures of proficiency in both languages. A comparison of the correlations with oral proficiency in the two languages, which is the only measure for which corresponding scores are available in the two languages, shows that the correlations between Accuracy and SPANSPK and between Accuracy and ENGSPK are equal for Form 1 but not equal for Form 2. For Form 2, the correlation between SEVTE Accuracy and ENGSPK was slightly higher.

---

[35] Accuracy requires the correct comprehension of the Spanish language propositions, whereas Expression does not. That is, one can score high on Expression and still not render an accurate translation.

[36] It is interesting to note that the self-ratings of translation ability, SPENSELF and ENSPSELF, also exhibit a similar difference in their correlations with SEVTE Accuracy and Expression, whereas the FBI's previous measure, SPENTRAN, does not exhibit any differential in the magnitude of its correlation with SEVTE Accuracy and Expression. This suggests that SPENTRAN seems to measure both constructs equally. On the other hand, ENSPTRAN does correlate more highly with SEVTE Accuracy than with Expression, suggesting that it focuses on accuracy, or that accuracy plays a more important role in the ENSPTRAN than in the SPENTRAN.

It was indicated earlier that Accuracy is the principal measure of translation ability while Expression focuses on the appropriateness of the usage in the target language document. Thus, one would expect higher correlations with the criterion variables for Accuracy than for Expression, which was also found to be true. The exception to this expectation would be the criterion variable that assesses English proficiency. Here, one would expect to find Expression correlating at least as high as Accuracy, and perhaps higher. An examination of the SEVTE Accuracy and Expression correlations with ENGSPK in Table 16 shows this expectation was met. Accuracy correlates .47 and .53 with ENGSPK and Expression correlates .50 and .51. Thus, the correlations with ENGSPK are equal.

7.4. Conclusions

From this discussion of the construct validity of the SEVTE through the examination of criterion-related, convergent and discriminant relationships with other measures, four conclusions can be reached.

First, SEVTE Accuracy and Expression measure different constructs. While the two constructs are correlated, the correlations (.74 to .75) are far from perfect. Thus, neither score can serve as a substitute for the other. The fact that a person can translate information accurately from Spanish does not mean that he or she can express it appropriately in English. Similarly, the fact that a person can express a translation appropriately in English does not mean that the information is accurate.

Second, both SEVTE Accuracy and SEVTE Expression appear to be valid measures. Both were found to correlate highly with translation skill levels assigned by comparing direct translations to the FBI/CAL translation skill level descriptions. SEVTE Accuracy was found to correlate with self-ratings of ability to translate various kinds of Spanish language documents on the job, with the FBI's current translation tests, with scores on all language proficiency tests, including measures of Spanish listening, speaking, and reading, and English speaking. Expression was found to correlate with all of the above measures, except Spanish speaking.

Third, Accuracy is the central construct. That is, Accuracy is the more valid measure of translation ability. In this study, Accuracy showed moderate to moderately high correlations with all the criterion variables. Expression is not as highly nor as consistently correlated with the criterion variables as Accuracy. Thus, Expression can be viewed to represent a secondary, although still important, construct in translation.

Fourth, an analysis of discriminant validity provides additional, generally positive evidence for the validity of both Accuracy and Expression. The SEVTE Accuracy measure correlates more highly with the FBI/CAL translation skill level for Accuracy than with the FBI/CAL translation skill level for Expression. The SEVTE Expression measure correlates more highly with the translation skill level for Expression than with the translation skill level for Accuracy. Both measures correlate more highly with self ratings of Spanish-English translation ability than with self ratings of English-Spanish translation ability. However, similarly clear evidence was not found in the correlations with the FBI's current tests of translation ability.

Finally, the SEVTE correlations with the various measures of language proficiency permit three additional conclusions about the role of various language skills in each SEVTE score.

First, English, the target language, plays a greater role in the Expression score than does Spanish, the source language. In this study, there was one measure of English proficiency and three measures of Spanish proficiency. The one English proficiency measure showed a greater correlation with Expression than did the Spanish measures.

Second, Spanish and English (the target and source languages) play approximately equal roles in the Accuracy score. In this study, all four language measures showed moderate to moderately high correlations with Accuracy. For the one skill where there were corresponding measures in both languages (speaking), the correlations were equal for Spanish and English on Form 1, but not equal for Form 2.

Third, Spanish, the source language, plays a greater role in the Accuracy score than in the Expression score. The data here showed that Spanish correlated higher with Accuracy than with Expression for the three skills measured (Spanish speaking, listening, and reading).

These conclusions about the role of proficiency in the two languages in the various scores provide additional insights into the skills required for Spanish into English translation.

8. Construction of Translation Skill Level Score Conversion Tables for the SEVTE

This section describes the construction of tables to convert raw scores on the SEVTE for Expression and Accuracy to FBI/CAL Translation Skill Levels (TSLs). In order to make decisions on the basis of test scores, compare test scores across forms, and interpret test scores, raw scores on the SEVTE must be converted to TSL scale scores.

8.1. Overview

In most of the preceding discussion of the SEVTE, raw scores have been used.[17] However, one of the goals of the project was to be able to interpret test scores in a way that is grounded in the Translation Skill Level Descriptions.[18] This entailed the construction of raw score-to-TSL score conversion tables for Expression and Accuracy for each section and each form of the test. These are presented in Appendix O.

Construction of the scaled score conversion tables is an attempt to give interpretative meaning to the SEVTE raw scores. In addition, it enables the comparison of total scores across forms and, to an extent, across the Multiple Choice section on the two forms. Conversion into scaled scores takes into account differences in test difficulty. Thus, a comparison of results across test forms and subtests must only be made in terms of the TSL scores.

---

[17] Weighted scores were used for many of the correlations involving Form 2 Expression scores.

[18] The Statement of Work in the RFP issued by the FBI for this project called for the development of a test "which would ultimately result in a score which can be converted to the 0 through 5 scale."

8.2. Determining Contributors to Expression and Accuracy Total Scores

Given the format of the test and the scoring system, there was a total of 185 possible points on the test when all the subscores were added together. However, after the data was collected, it became apparent that there should be separate scores for Expression and Accuracy. (See the discussion of the the history of the SLDs and the discussion of the constructs in sections 1.4.1. and 1.5.3.) Based on our conceptualization of the constructs, it was clear that scores for paragraph expression (PEX), paragraph grammar (PGR) and paragraph mechanics (PME) should contribute to the total Expression score, while sentence accuracy (SAC) and paragraph accuracy (PAC) should contribute to the total Accuracy score. To determine to which score the Multiple Choice (MC) section and the Word and Phrase Translation subsection belonged, a multiple-regression "r-square" analysis was performed. An r-square analysis determines the r-square value (percent of variance shared by the combination of the variables with the criterion) of all combinations of the variables entered into the equation when regressed on the criterion (overall EXPFBICAL and overall ACCFBICAL). Both MC scores and Word and Phrase Translation scores were entered into the r-square analysis together with PEX, PGR and PME, using the overall FBI/CAL Expression score as a criterion. In addition, both MC scores and Word and Phrase Translation scores were entered into the r-square analysis together with SAC and PAC, using the overall FBI/CAL Accuracy score as a criterion. The results of all the r-square analyses (Expression and Accuracy scores for the two forms of the SEVTE and the two forms of the ESVTE) were examined together. The results indicated that, although MC and Word and Phrase Translation scores contributed to both Expression and Accuracy scores, the most parsimonious combination of scores was for MC to be used as a subscore for Expression and Word and Phrase Translation as a subscore for Accuracy.

Once these combinations of subscores were determined, we examined whether there was anything to be gained by differentially weighting the different subscores to produce the total score. Regressions were run to determine the maximum amount of variance shared between the optimal combination of subscores and the corresponding criterion variable. These were compared to forming total scores without differential weighting. This analysis revealed that little was to be gained by weighting in all cases except the total Expression score for Form 2 of the SEVTE. The correlation with the FBI\CAL translation skill level rating for Expression were significantly improved by the assignment of different weights to the Form 2 Expression subsections. Thus, the weights for Form 2 Expression were set as follows:

Total Form 2 Expression = .289 x Form 2 MC +
1.920 x Form 2 PGR +
.456 x Form 2 PME +
3.466 x Form 2 PEX.

This combination of weights indicates that paragraph expression and paragraph grammar receive greater emphasis while paragraph mechanics and the total multiple choice section scores receive lesser emphasis than in the Form 1 total Expression score, which is scored solely on the basis of raw score points. SEVTE Form 2 was the only one of the six test forms developed as part of this project that profited significantly from differential weighting.

8.3. Development of Raw Score to Scaled Score Conversion Tables

Since one of the goals of the project was to provide translation ability scores based on the TSL descriptions, it was necessary to identify a procedure that would anchor SEVTE scores, which are analytical, to the holistic TSL descriptions. This was accomplished during the validation study (see section 7.2) by having each rater assign to each paper, separately for Expression and Accuracy, a translation proficiency skill level based on the FBI/CAL translation skill level descriptions. This procedure produced in four holistic ratings for Accuracy and four holistic proficiency ratings for Expression. These two sets of four holistic proficiency ratings were then averaged separately, to give each examinee an overall FBI/CAL TSL score for Expression and Accuracy.

To develop a conversion table of raw SEVTE scores to TSL scores, total raw scores for Expression and Accuracy for all subjects were averaged between raters, with the Expression score for Form 2 being weighted. These total raw scores were then regressed on the corresponding overall FBI\CAL translation skill level (Expression or Accuracy). As shown in Table 15, correlations between the total SEVTE scores and these overall scores were very high: from .88 to .89 for Expression and from .89 to .92 for Accuracy. These high correlations produced optimal regression equations for predicting TSL scores from raw scores on each form of the test. These equations were then used to produce predicted TSL scores from all possible SEVTE scores for each form.[39] These conversion tables are presented in Appendix O.

8.4. Using the Multiple Choice Section as a "Screen"

The Multiple Choice section of the SEVTE may be used to screen out individuals for whom the production section of the test is inappropriate. Section 2.4 of this report describes how it was determined to use the multiple choice section score as a screen. The Multiple Choice score selected (mentioned below) is the best predictor of a TSL rating of 2.0 on the combined multiple-choice and production sections of the SEVTE. Examinees who score below this level are unlikely to score a 2.8 (2+) or above on the total test after their raw score has been converted to the corresponding TSL score for Accuracy. The SEVTE total score corresponding to a TSL of 2+ is the recommended passing score; that is, the score at which examinees can serve as translators for the FBI.

---

[39] For a considerable number of examinees on each form of the test, this regression line resulted in a perfect prediction. That is, the overall TSL rating predicted by applying the regression line to the raw score (or weighted score in the case of Form 2 Expression) coincided exactly with the average TSL rating assigned by the rater. However, there was a tendency toward greater error among examinees who scored higher on the SEVTE. This was due to a number of causes, including the regression effect, sampling, and the speededness of the Paragraph Translation subsection during the validation study. For additional information on the accuracy of predicted Translation Skill Levels see CAL's memo to the FBI dated May 15, 1990, in Appendix P.

In using the SEVTE MC as a screen, the most serious error one can make is to exclude someone from taking the Production section who may ultimately score a 2+ or above. Giving the Production section to someone who may not ultimately score 2+ or above is not a serious error, since this individual will ultimately be evaluated correctly (after the production section is scored). To determine the cut-off score on the Multiple Choice section, we need to determine the raw score on the Multiple Choice section that corresponds to a TSL score of 2; that is, we need to determine the raw score on the MC section that corresponds to a translation proficiency level of 2 for Accuracy.

To determine the raw score on the MC section that corresponds to a score of 2, raw scores on the MC section were regressed on the overall Accuracy scores. (Note that for Form 1 the correlation between these two scores was .76; for Form 2 it was .69. The root mean square error of the regression for Form 1 was .470 of a level; for Form 2 it was .492.) This analyses revealed that the score of 33 would be the lowest predictor of a score in the 2 range on Form 1, while 25 would be the lowest predictor of that score for the more difficult Form 2. These, then, are the recommended cut-off scores on the Multiple Choice section. Examinees who score below this level on the Multiple Choice section of the SEVTE either need not take the production section, or if they already have, that section need not be scored.

Using these cut-off scores would still leave in many examinees who may not ultimately achieve a score at or above 2+ in Accuracy on their total test; however, the probability of excluding a candidate who might achieve a 2+ in Accuracy on the total test is minimal.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Brennan, R.L. (1983). *Elements of generalizeability theory*. Iowa City, IA: The American College Testing Program.

Center for Applied Linguistics. (September 8, 1987). *Proposal to develop Spanish-English English-Spanish translation tests.* Washington, DC: Center for Applied Linguistics.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Duran, R.P., Canale, M., Penfield, J., Stansfield, C.W. & Liskin-Gasparro, J.E. (1985). *TOEFL from a communicative viewpoint on language proficiency: A working paper*. Princeton, NJ: Educational Testing Service. Alexandria, VA: ERIC Document Reproduction Service No. ED 263 127.

Federal Bureau of Investigation. (August 7, 1987). *Request for Proposals No. 4327*. Washington, DC: Federal Bureau of Investigation.

Kachru, B.J. (1985). Standards, codification, and sociolinguistic realism: The English language in the outer circle. In R. Quirk and H.G. Widdowson, (eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11-30). Cambridge: Cambridge University Press.

Newmark. P. (1981). *Approaches to translation*. Oxford: Pergamon Press.

Pochhacker, F. (1989). Beyond equivalence: recent developments in translation theory. In D.L. Hammond, Ed., *Coming of age*. Proceedings of the 30th Annual Conference of the American Translators Association (pp. 563-571). Medford, NJ: Learned Information Inc.

Stansfield, C.W., Scott, M.L., and Kenyon, D.M. (1990). *Listening Summary Translation Exam (LSTE) - Spanish*. Final project report: revised. Washington, DC: Center for Applied Linguistics.

Stansfield, C.W., Scott, M.L., & Kenyon, D.M. (1990). English - Spanish Verbatim Translation Exam. Final report. Washington, DC: Center for Applied Linguistics.

Walker, M., Williams, M., & Navarrete, O. (1988). Aptitude and language learning of FBI Special Agents. Paper presented at the ILR Invitational Symposium on Language Aptitude Testing, Arlington, VA. Alexandria, VA: ERIC Document Reproduction Service No. ED 307 797.

APPENDIX A

ADMINISTRATION INSTRUCTIONS FOR SEVTE

TEST ADMINISTRATION INSTRUCTIONS

SPANISH INTO ENGLISH VERBATIM TRANSLATION EXAM

DEPARTMENT OF JUSTICE
FEDERAL BUREAU OF INVESTIGATION

NOTE TO TEST ADMINISTRATOR

This manual describes important information about the procedures that must be followed BEFORE, DURING, and AFTER the administration of the translation exams. Uniform procedures are essential for the translation exams to yield reliable test results. The scores of all examinees from various field offices in the nation will be comparable only if all test administrators follow the same procedures and give exactly the same instructions. It is necessary, therefore, that you read the entire manual before administering the exams and follow the instructions without exception when administering the exams.

GENERAL INFORMATION

Test Security

It is extremely important that the translation exams be safeguarded and administered under secure conditions at each field office. In order to ensure test security, it is essential that you adhere to the following conditions:

1. Keep all test materials either in your immediate physical possession or in a locked cabinet or other secure area under your control.
2. Do not copy, or allow others to copy, any portion of the test booklets or tape, or make any notes or transcriptions of the test booklets or tape content.
3. Allow only those particular individuals who are to be tested to see the test materials, and only at the time of test administration and under the specific procedures described in this manual.
4. Should any irregularities occur, report them on the Test Administrator Report Form included in the test package. Please complete and sign this form even if no irregularities occur.

PRIOR TO THE TESTING DATE

Assembling Test Materials

Assemble as many test booklets and answer sheets as will be needed for the test administration, including two or three extra copies of each. You should also have on hand at least two no. 2 pencils (with erasers) for each examinee. Listed below are the materials needed for each exam:

1) Multiple Choice Section test booklets
2) Production Section test booklets
3) Answer sheets
4) No. 2 pencils
5) A timer, wristwatch or other timepiece which can be reset

Arranging for a Testing Site

Locate a testing site that is comfortable and free from distraction. The testing room should be large enough so that examinees can be seated with three feet of space in all directions between all examinees.

ON THE TESTING DATE

Equipment

Check to make sure the timepiece is functioning properly and has been completely reset to zero (or 12:00). There should always be at least two timepieces in the testing room as a check against mistiming.

Prohibited Materials

While taking the Multiple Choice Section and the Translation of Words and Phrases in Context and Sentence Translation Section, examinees should not have anything on their desks except their pencils, test booklets, and answer sheets. Examinees may use dictionaries only during the Paragraph Translation Section.

Administering the Test

Follow the procedures below when administering the test. All instructions within the boxes should be read verbatim. Pause where four dots appear to allow time for the procedure described to be carried out. Be sure you state the correct form where appropriate. Do not depart from these directions unless noted otherwise.

1. After all examinees have been seated, distribute the Multiple Choice Section test booklets, answer sheets, and pencils.

2. Give the following instructions:

> **Please do not open your test booklet. In this section of the exam, you will mark all of your answers on the answer sheet. Do not write anything in the test booklet. You must use a no. 2 pencil for marking your answers.**

3. Instruct the examinees how to fill out the answer sheet:

Place your answer sheet on top of your test booklet. Turn the answer sheet so that you see SIDE ONE in the upper right hand corner...

On the left half of side one, you will see an area containing blue lines. At the top of this section is the word NAME. Print your name in the boxes provided. Print your last name, and then your first name. Leave a blank space between your last name and your first name...

Now fill in the circles beneath the boxes in which you printed your name. Each circle you fill in must correspond to the letter you printed in the box above. Be sure that you darken the circle so that the letter within the circle is completely covered. You should not be able to see the letter. If you make a mistake, erase the mistake completely. Do not make any extra marks on your answer sheet. Your answer sheet will be scored by a machine. If you do not mark it carefully, it may not be processed accurately by the scoring machine.

Now find the section labeled IDENTIFICATION NUMBER in the bottom left half of your answer sheet. Print your SOCIAL SECURITY NUMBER in the boxes labeled A through I...

Now fill in the circles beneath the boxes in which you printed your social security number. Each circle you fill in must correspond to the number you printed in the box above...

Now find the section labeled SPECIAL CODES, located to the right of the section you just completed. [GIVE THE FOLLOWING INSTRUCTIONS IN ACCORDANCE WITH THE FORM NUMBER OF THE EXAM YOU ARE NOW ADMINISTERING:] Print the number [ONE or TWO] in box K. This is [FORM 1 or FORM 2] of the Spanish into English Verbatim Translation exam. You do not need to fill in your birth date, sex, or level of education...

Now look at the right half of your answer sheet. Notice that the first fifty items are arranged in columns in the top section of the answer sheet, while the next fifty items are arranged in the bottom section. Make sure you follow the order of the items as they are marked. For example, after question number ten, you will need to return to the top of the section to mark your answer to question number eleven.

Are there any questions?_Try to answer every item, but do not be concerned if you can not answer all of them. You will not be penalized for guessing. If you are unsure of the answer to a question, make the best guess you can and go on to the next question. The verbatim translation exam takes approximately two hours and ten minutes to complete.

4. Instruct the examinees to begin the Multiple Choice Section:

5. Walk about the room to make sure that everyone is marking their answers correctly on the answer sheet.

Now remove from your desk everything except your test booklet, answer sheet, pencils, and erasers_

Look at your test booklet for the Multiple Choice Section of the Spanish into English Verbatim Translation Exam. Print your name in the space provided on the cover. Print your last name first_

Print today's date in the space provided_

There are two parts in this section. You will be allowed a total of thirty-five minutes to complete both parts. I will advise you when there are five minutes remaining. You may now open your test booklets and begin the test. [START TIMER IMMEDIATELY]

6. After 30 minutes, inform examines:

There are five minutes remaining to complete this section.

7. After 35 minutes, STOP AND RESET THE TIMER. Inform examinees:

> **This is the end of the Multiple Choice Section. Please stop working now. Now look over your answer sheet carefully. Be sure all the marks you made are dark and heavy. Insert your answer sheet in your test booklet and close the booklet.**

8. Collect the test booklets and answer sheets for the Multiple Choice Section. Be sure to account for all test booklets distributed.

9. Distribute the Words and Phrases in Context and Sentence Section booklets. Instruct the examinees to begin this section:

> **There are two parts in the next section. You may not use your dictionary during this section. You will be given 35 minutes to complete the two parts in this section, the Translation of Words and Phrases in Context and Sentence Translation. I will advise you when there are five minutes remaining to finish this section. You may now open your test booklets and begin working. [START TIMER IMMEDIATELY]**

10. After 30 minutes, inform examinees:

> **There are five minutes remaining to complete this section.**

11. After 35 minutes, STOP AND RESET THE TIMER. Inform examinees:

> **Please stop working now. We will now have a short rest break. We will begin the Paragraph Translation Section in five minutes. You may leave the room if you wish.**

12. Collect the test booklets for the Words and Phrases in Context and Sentence Section. Be sure to account for all test booklets distributed.

13. Distribute the Paragraph Translation Section booklets. Instruct the examinees to begin the Paragraph Translation Section:

> **We will now begin the Paragraph Translation Section. In this section you will translate three paragraphs. You may use dictionaries during this part of the exam. You will have 48 minutes to complete the Paragraph Translation Section. I will inform you when there are five minutes remaining. When you have finished this section, please close your test booklets and wait for further instructions. You may now begin. [START TIMER IMMEDIATELY]**

14. After 43 minutes, inform examinees:

> **There are five minutes remaining.**

15. After 5 minutes, inform examinees:

> **Please stop working now. Close your test booklets.**

16. Collect the test booklets for the Paragraph Translation Section.

Test Administrator Report Form

SPANISH INTO ENGLISH VERBATIM TRANSLATION EXAM

This form is to be used to report any irregularities in test administration. Please fill it out (even if there were no irregularities), sign your name, and return it with the test materials. Thank you.

• • • • • • • • •

Test Security

By agreeing to serve as the test administrator, I am responsible for ensuring the security of the test. I have kept the test materials confidential and secure at all times. None of the test booklets or test tapes has been reproduced in any form.

Irregularities: ______________________________

____

Test Administration

The tests were administered in exact accordance with the procedures described in the Administration Manual. Any deviations from the stated procedures are listed below:

Irregularities: ______________________________

____

Condition of Test Materials

Before returning the test materials, I have checked the condition of the test booklets and test tapes. All materials are being returned in their original condition.

Irregularities: ______________________________

____

____

(Please print name) Field Office

____

Signature Date

APPENDIX B

MULTIPLE CHOICE SECTION TITLE PAGE AND INSTRUCTIONS

NAME ______________________________
Last First

DATE ________

SPANISH INTO ENGLISH VERBATIM TRANSLATION EXAM

MULTIPLE CHOICE SECTION

FORM 1

DEPARTMENT OF JUSTICE
FEDERAL BUREAU OF INVESTIGATION

This test is for official use only; do not divulge any information contained herein. Do not duplicate any portion of this test. Do not show to unauthorized persons.

FIELD OFFICE ______________________________

TEST NO. _____

SPANISH INTO ENGLISH VERBATIM TRANSLATION EXAM (SEVTE)

MULTIPLE CHOICE SECTION: INSTRUCTIONS AND EXAMPLE ITEMS

EMBEDDED PHRASE ITEMS

Instructions: Choose the best translation for the underlined portions of the following sentences. If there is more than one possible answer, choose the most appropriate translation. Consider how the entire sentence should be translated when choosing the correct answer. On your answer sheet, find the number of the question and blacken the space that corresponds to the letter of the answer you have chosen.

Example: Dicen que mañana va a llover.

(A) to snow
(B) to cry
(C) to rain
(D) to call

Discussion: The translation of the full sentence is, They say that tomorrow it's going to rain. To rain is the correct translation of llover; therefore, the answer is (C). You would black the space marked (C) on your answer sheet.

ERROR DETECTION ITEMS

Instructions: Blacken the space corresponding to the letter of the incorrect part of the sentence on your answer sheet. If there is no error, choose (D). There cannot be more than one error in each sentence. Possible errors include: incorrect grammar, word order, vocabulary, punctuation or spelling.

Example: You shouldnt forget to call her tomorrow.
A B C

Discussion: The apostrophe has been omitted from the contraction shouldn't; therefore, the correct choice is (A). You would blacken the space marked (A) on your answer sheet.

APPENDIX C

PRODUCTION SECTION TITLE PAGE AND TEST INSTRUCTIONS

NAME ______________________________
Last   First

DATE __________

SPANISH INTO ENGLISH VERBATIM TRANSLATION EXAM

PRODUCTION SECTION

FORM 1

DEPARTMENT OF JUSTICE
FEDERAL BUREAU OF INVESTIGATION

This test is for official use only; do not divulge any information contained herein. Do not duplicate any portion of this test. Do not sho- unauthorized persons.

FIELD OFFICE ______________________________

TEST NO. _____

SPANISH INTO ENGLISH VERBATIM TRANSLATION EXAM (SEVTE)

PRODUCTION SECTION: INSTRUCTIONS AND EXAMPLE ITEMS

TRANSLATION OF WORDS AND PHRASES IN CONTEXT

Instructions: After you have read each of the following sentences, translate the underlined portion into English. Strive for a natural, grammatical rendition which doesn't modify the original meaning. Consider how the entire sentence would be translated before providing your answer. Use the spaces below each sentence.

Example: Les he contado mucho de tí a mis padres.

I have told

Discussion: In this case, the pronoun les is not translated because the meaning is already contained in the translation of the full noun phrase of the indirect object: my parents. The translation of the complete sentence would be: **I have told my parents a lot about you.** It would not be correct in English to use both the pronoun them and the noun phrase my parents in this sentence.

SENTENCE TRANSLATION

Instructions: After you have read the following sentences, translate them into English. Use the spaces provided. Make sure your rendition sounds natural in English while retaining the original meaning.

Example: Los países en vías de desarrollo necesitan la ayuda de las naciones industrializadas.

Developing countries need the assistance of industrialized nations.

Discussion: Note that developing countries is an appropriate translation of the idiomatic expression países en vías de desarrollo. A more literal translation (i.e., countries on the road to development) would not sound natural in English. Note also that the definite article the is not used in the English translation of either plural noun phrase (i.e., developing nations and industrialized nations). Additionally, the placement of the adjective industrialized is in front of the noun in English.

PARAGRAPH TRANSLATION

Instructions: Translate the following paragraphs into English. Again, strive for a natural rendition without changing the original meaning. You are permitted to use a dictionary during this section only. Do not return to work on previous sections.

APPENDIX D

CONTENT ANALYSIS OF SEVTE MULTIPLE CHOICE SECTIONS

CONTENT ANALYSIS

SPANISH-ENGLISH (EXAM I)

1. vocabulary - adverbial phrase
2. vocabulary - idiom (complete phrase)
3. vocabulary - adverb
4. grammar - use of subjunctive
5. vocabulary - conjunction
6. vocabulary - verb phrase
7. vocabulary - adverbial phrase
8. vocabulary - adverbial phrase
9. vocabulary - verb phrase
10. vocabulary - false cognate (verb)
11. a. vocabulary - verb phrase
    b. grammar - use of subjunctive
12. vocabulary - false cognate (verb)
13. vocabulary - false cognate (verb)
14. vocabulary - verb phrase
15. vocabulary - false cognate (adjective)
16. a. vocabulary - verb
    b. grammar - use of subjunctive
17. vocabulary - adverb
18. vocabulary - false cognate (adverb)
19. vocabulary - adverbial phrase
20. vocabulary - noun phrase
21. vocabulary - verb phrase
22. vocabulary - noun
23. vocabulary - adjective
24. vocabulary - false cognate (noun phrase)
25. vocabulary - false cognate (noun)
26. vocabulary - proverb
27. vocabulary - false cognate (verb)
28. vocabulary - noun
29. grammar - use of subjunctive
30. vocabulary - verb phrase
31. vocabulary - verb phrase
32. vocabulary - verb phrase
33. vocabulary - verb phrase
34. vocabulary - noun phrase
35. vocabulary - verb
36. spelling
37. grammar - verb form (past participle)
38. grammar - subject-verb agreement
39. grammar - verb form
40. grammar - verb form
41. spelling
42. grammar - use of pronoun (subject-verb agreement with pronoun)
43. vocabulary - false cognate (adjective)
44. No error
45. vocabulary - false cognate (noun)

46. grammar - subject-verb agreement
47. No error
48. grammar - verb form (use of infinitive vs. present participle)
49. punctuation - use of apostrophe
50. punctuation - comma
51. No error
52. spelling
53. grammar - use of pronoun (inconsistency)
54. grammar - use of pronoun (pronoun-noun agreement)
55. grammar - use of pronoun (subjective vs. objective)
56. grammar - use of pronoun (relative - who/whom)
57. grammar - adjective-noun agreement (less/fewer)
58. grammar - use of pronoun (relative - who/which)
59. vocabulary - conjunction
60. grammar - (lie/lay)

GRAMMAR is tested:	18 times	
verb form:	4 times	
use of subjunctive:	4 times	
subject/verb agreement:	2 times	
use of pronouns:	6 times	
adjective/noun agreemtent:	1 time	
lie vs. lay	1 time	
VOCABULARY is tested:	36 times	
adjective or adjectival phrase:	3 times	(2 FC)
adverb or adverbial phrase:	7 times	(1 FC)
noun or noun phrase:	7 times	(3 FC)
verb or verb phrase:	15 times	(4 FC)
proverb:	1 time	
conjunction:	2 times	
idiom:	1 time	
PUNCTUATION is tested:	2 times	
SPELLING is tested:	3 times	
NO ERROR appears:	3 times	

CONTENT ANALYSIS

SPANISH-ENGLISH (EXAM II)

1. vocabulary - adverbial phrase
2. vocabulary - idiom (complete phrase)
3. vocabulary - adverbial phrase
4. grammar - use of subjunctive
5. vocabulary - conjunction
6. vocabulary - verb phrase
7. vocabulary - adverbial phrase
8. vocabulary - adverbial phrase
9. vocabulary - adverbial phrase
10. vocabulary - false cognate (verb)
11. a. vocabulary - verb
    b. grammar - use of subjunctive
12. vocabulary - verb
13. a. vocabulary - false cognate (verb)
    b. grammar - use of preposition
14. vocabulary - verb phrase
15. vocabulary - adjective phrase
16. a. vocabulary - verb
    b. grammar - use of subjunctive
17. vocabulary - adverb
18. vocabulary - false cognate (noun)
19. vocabulary - adverb phrase
20. vocabulary - noun phrase
21. vocabulary - verb phrase
22. vocabulary - noun
23. vocabulary - adjective
24. vocabulary - verb phrase
25. vocabulary - false cognate (noun phrase)
26. vocabulary - proverb
27. vocabulary - false cognate (verb phrase)
28. vocabulary - idiom (complete phrase)
29. grammar - use of subjunctive
30. vocabulary - verb phrase
31. vocabulary - verb phrase
32. vocabulary - verb phrase
33. vocabulary - verb phrase
34. vocabulary - noun phrase
35. vocabulary - verb
36. spelling
37. grammar - past participle
38. grammar - subject-verb agreement
39. grammar - verb form
40. grammar - verb form
41. spelling
42. grammar - subject-verb agreement with pronoun
43. vocabulary - false cognate (noun)
44. No error
45. vocabulary - false cognate (noun)

46. grammar - subject-verb agreement
47. No error
48. grammar - verb form (infinitive vs. present participle)
49. punctuation - use of apostrophe
50. punctuation - comma
51. No error
52. spelling
53. grammar - use of pronoun (inconsistency)
54. grammar - use of pronoun (pronoun-noun agreement)
55. grammar - use of pronoun (subjective-objective)
56. grammar - use of pronoun (relative - who/whom)
57. grammar - noun-adjective agreement (less/fewer)
58. grammar - use of pronoun (relative - who/which)
59. vocabulary - conjunction
60. grammar - lie vs. lay

GRAMMAR is tested:	19 times	
verb form:	4 times	
use of subjunctive:	4 times	
subject/verb agreement:	2 times	
use of pronouns:	6 times	
adjective/noun agreemtent:	1 time	
lie vs. lay	1 time	
use of prepositions:	1 time	
VOCABULARY is tested:	36 times	
adjective or adjectival phrase:	2 times	
adverb or adverbial phrase:	7 times	
noun or noun phrase:	7 times	(4 FC)
verb or verb phrase:	15 times	(3 FC)
proverb:	1 time	
conjunction:	2 times	
idiom:	2 times	
PUNCTUATION is tested:	2 times	
SPELLING is tested:	3 times	
NO ERROR appears:	3 times	

APPENDIX E

SENTENCE ACCURACY SCORING GUIDELINES

FINAL VERSION

SENTENCE ACCURACY SCORING GUIDELINES

0 Translation is less than 50% complete.
1 Many mistranslations, omissions, and/or inappropriate additions, so that much of the meaning is lost.
2 Mistranslation or omission of one or more key terms (including verb tense), and/or inappropriate additions.
3 Mistranslation or omission of one or more minor terms; no inappropriate additions.
4 No mistranslations or omissions, although some nuance may not be conveyed.
5 All nuances conveyed.

APPENDIX F

PARAGRAPH SCORING GUIDELINES

FINAL VERSION

SEVTE PARAGRAPH SCORING GUIDELINES

GRAMMAR* (Structure and Morphology)

0 (Translation less than 50% complete.)
1 Majority of structures are incorrect.
2 Some errors in basic structures and numerous errors in complex structures.
3 Errors in basic structures are rare. Sporadic errors in high frequency complex structures; some errors in low frequency complex structures.
4 No more than one error in a complex structure.
5 No grammar errors.

EXPRESSION (Word Order, Vocabulary, Idiomaticity, Style, and Tone)

0 (Translation less than 50% complete.)
1 Expression generally equivalent to source language; unacceptable in target language.
2 Expression closer to source language; generally unacceptable in target language.
3 Expression usually follows target language conventions, but is not always preferred.
4 Expression occasionally reveals translation. Appropriate register.
5 No evidence of translation.

MECHANICS (Spelling. Punctuation, and Capitalization)

0 (Translation less than 50% complete.)
1 Numerous errors in spelling or punctuation.
2 Frequent errors in spelling or punctuation.
3 Occasional errors in spelling or punctuation.
4 Rarely makes errors in spelling or punctuation.
5 Almost no errors in spelling or punctuation.

ACCURACY

0 (Translation less than 50% complete or less than 50% accurate.)
1 Many mistranslations, omissions, and/or inappropriate additions, so that much of the meaning is lost.
2 Mistranslation or omission of one or more key terms (including verb tense), and/or inappropriate additions.
3 Mistranslation or omission of one or more minor terms; no inappropriate additions.
4 No mistranslations or omissions, although some nuance may not be conveyed.
5 All nuances conveyed.

* Use the information on the following page as a guide in distinguishing errors in basic, high frequency complex, and low frequency complex structures.

1) BASIC STRUCTURES: (subject/verb agreement, number [plural, singular], present tense, present progressive, simple past, pronouns, comparatives, going to future, 's possessives, present tense modals [can, will, shall, may, might, must])

2) HIGH FREQUENCY COMPLEX STRUCTURES: (articles, present perfect, past perfect, past progressive, past modals (could, would), perfect modals [must, could, might, may + have], used to, derivational endings [noun, adjective, adverb, verb endings], relative clause pronouns, tense sequencing, prepositions)

3) LOW FREQUENCY COMPLEX STRUCTURES: (gerunds vs. infinitives, subjunctive, conditional tense, future perfect, compound tenses [past perfect progressive, future perfect progressive, etc.], two word verbs [take over, take on, take up, etc.])

APPENDIX G

PILOT VERSION OF SENTENCE SCORING GRID

SENTENCE SCORING GRID

GRAMMAR

0 Less than 50% complete.
1 One or more errors in basic structures.
2 One or more errors in high frequency complex structures.
3 One or more errors in low frequency complex structures.
4 One error in a very low frequency complex structure.
5 No errors.

EXPRESSION

0 Less than 50% complete.
1 Expression generally equivalent to source language; unacceptable in target language.
2 Expression closer to source language; generally unacceptable in target language.
3 Expression follows target language conventions, but is not preferred.
4 Expression gives subtle indication of translation. Appropriate register.
5 No evidence of translation.

MECHANICS

0 Less than 50% complete
1 Four errors
2 Three errors
3 Two errors
4 One error
5 No error

ACCURACY

0 Less than 50% complete.
1 Many mistranslations, omissions, and/or inappropriate additions.
2 Mistranslation or omission of one or more key terms (including verb tense), and/or inappropriate additions.
3 Mistranslation or omission of one or more minor terms; no inappropriate additions.
4 No mistranslations or omissions, although some nuance may not be conveyed.
5 All nuances conveyed.

APPENDIX H

PILOT VERSION OF PARAGRAPH SCORING GRID

PARAGRAPH SCORING GRID

GRAMMAR

0 Less than 50% complete.
1 Majority of structures are incorrect.
2 Some errors in basic structures and numerous errors in complex structures.
3 Errors in basic structures are rare. Sporadic errors in high frequency complex structures; some errors in low frequency complex structures.
4 No more than one error in a low frequency complex structure.
5 No grammar errors.

EXPRESSION

0 Less than 50% complete.
1 Expression generally equivalent to source language; unacceptable in target language.
2 Expression closer to source language; generally unacceptable in target language.
3 Expression usually follows target language conventions, but is not always preferred.
4 Expression occasionally reveals translation. Appropriate register.
5 No evidence of translation.

MECHANICS

0 Less than 50% complete
1 At least 50% correct
2 At least 70% correct
3 At least 80% correct
4 At least 90% correct
5 At least 99% correct

ACCURACY

0 Less than 50% complete.
1 Many mistranslations, omissions, and/or inappropriate additions.
2 Mistranslation or omission of one or more key terms (including verb tense), and/or inappropriate additions.
3 Mistranslation or omission of one or more minor terms; no inappropriate additions
4 No mistranslations or omissions, although some nuance may not be conveyed.
5 All nuances conveyed

APPENDIX I

FBI/CAL TRANSLATION SKILL LEVEL DESCRIPTIONS AND QUESTIONNAIRE

July 26, 1990

FBI/CAL TRANSLATION SKILL LEVEL DESCRIPTIONS

EXPRESSION

0+ Makes very frequent mistakes in spelling, punctuation, and representation of symbols. Uses none or almost none of the morphology or syntax conventions of the target language. Vocabulary is extremely limited and frequently inappropriate, even when using a dictionary. Only very simple sentences are correct. Style and tone are not identifiable. Renders a translation that appears very distorted and for the most part is unintelligible.

1 Makes frequent spelling and punctuation errors, frequent grammar errors in basic structures, and shows little ability to convey verb tenses other than the present tense. Syntax is generally equivalent to that of source language. Vocabulary is often inappropriate, even when using a dictionary, and active vocabulary is usually limited to everyday words and cognates. Renders an extremely literal translation, i.e. almost word by word. Has no ability to deal with complex sentence patterns. Unable to convey style and tone, unless their use in source document is very predictable. Portions of the translation are unintelligible and others are clearly distorted; however, much of it can be understood by native readers used to dealing with foreigners' efforts to translate their language.

1+ Makes many spelling errors and punctuates according to source language conventions. Makes many errors in basic grammatical structures, and uses very few low frequency constructions correctly. Uses syntax that is very close to that of source language, while vocabulary is limited and makes many errors in choice of words, sometimes even when using a dictionary. Attempts at complex sentences often result in errors. Uses uneven style and tone that do not reflect those of original document. This person's translated documents appear distorted but are mostly intelligible to native readers used to dealing with foreigners' efforts to translate their language.

2 Makes spelling errors, while capitalization and punctuation errors reflect source language conventions. Uses syntax that is closer to source language than to target language. Makes very frequent errors in low frequency grammatical structures, frequent errors in high frequency grammatical structures, and some errors in basic structures. Vocabulary may be generally too limited to convey abstract thoughts. Has only some knowledge of idiomatic expressions and colloquialisms, and very limited knowledge of sayings and proverbs. Distorts the style and/or the tone of the original document and may inappropriately combine use of formal and informal patterns of speech. Produces translations that are very literal, but are generally understandable to a native reader NOT used to dealing with foreigners' efforts to translate their language.

2+ Makes some spelling errors, and may use capitalization and punctuation that imitates usage of source language. Uses syntax that tends to reflect that of source language. May make frequent errors in low frequency complex grammatical structures, some errors in high frequency complex structures, and occasional errors in basic structures. Has little ability to use complex sentence patterns. Vocabulary is adequate to express some abstract thoughts; can often make sensible guesses about unfamiliar words using linguistic context and prior knowledge. Has a fair knowledge of idiomatic expressions and colloquialisms and only limited knowledge of sayings and proverbs. Tone and style are uneven and somewhat distorted. Produces documents that are readily understandable but clearly have been translated.

3 Occasionally makes spelling mistakes, some grammar mistakes in low frequency complex structures, sporadic errors in high frequency complex structures, and shows no pattern of errors in basic structure. Uses punctuation that is almost identical to source document, i.e. sometimes atypical of the target language. Moderately good ability to join or divide original sentences as required by target language constructions, while still retaining the meaning of the source document. Moderately good ability to use complex structures, sentence patterns, and vocabulary appropriate for expressing abstract thoughts. Moderately good knowledge of idiomatic expressions and colloquialisms, and some sayings and proverbs, but with occasional misunderstandings. Uses a number of syntactic constructions that are more characteristic of source language than target language, thereby producing documents that appear to be a translation. This person's style and tone are even, but occasionally differ slightly from original.

3+ Makes occasional spelling and punctuation errors. Occasionally makes grammatical errors in low frequency complex structures, sporadic errors in high frequency complex structures. Good ability to use very complex sentence structures. Uses some syntactic structures that are more typical of source than target language which suggest that the document is translated. Vocabulary is generally extensive but usage is not always precise given the context, especially in the use of register and colloquialisms. The style and tone of the original document are not always retained.

4 This person's errors of grammar are very rare and unpatterned. This person rarely makes a spelling or punctuation error. Uses some syntactic structures that suggest the document is a translation--while these are grammatically correct, they are not typical of the target language. Very good ability to use highly complex sentence structures. Very good knowledge of idiomatic expressions, register, colloquialisms, sayings and proverbs and their equivalents in the target language. However, a document rendered by this person may occasionally reveal itself to be a translation due to atypical use of syntax and vocabulary. The style and tone are equivalent to those of the source document.

4+ Makes no grammatical or punctuation errors, and no spelling errors that would not be made by an educated native writer of the target language. There are minor problems of syntax, spelling, or vocabulary, which although grammatically correct are not typical of the source language and suggest that the document is a translation. These and other infelicities could only be confirmed by an educated native reader of both languages who compares the documents in both the source language and the target language. Uses style and tone that are a true reflection of source document.

5 Produces work that contains no grammar, spelling or punctuation errors that would not be made by other well-educated native writers. Can produce documents whose syntax is that of the target language, with no influence of source language. Can adapt rhetorical structures so that the document reads as if it had originally been written in the target language. Can convey all nuances and can use tone and stylistic devices that are identical in effect to those of original, including use of humor.

ACCURACY

0+ Has no real ability to translate connected discourse. Efforts to translate contain many mistranslations and omissions, and very little information from source document is conveyed.

1 Renders translations whose accuracy is deficient, with frequent mistranslations and omissions and may make inappropriate additions. Much of the information from longer source documents is lost.

1+ Produces translations whose accuracy is inadequate, containing many mistranslations or omissions, and possibly additions. Almost all nuances are lost.

2 Produces translations whose accuracy is mostly adequate and without severe substantive omissions, but without many nuances, and with quite a few mistranslations. May include some additions for clarification of areas the translator can not accurately convey.

2+ Produces translations whose accuracy is adequate, but contain some mistranslations or omissions, and reflect a limited ability to convey nuances.

3 Produces translations whose accuracy is good, with occasional minor mistranslations or omissions. Can handle clearly identifiable nuances.

3+ Produces translations whose accuracy is very good; there are occasional omissions, or sporadic minor mistranslations; nuances and subtleties are not always conveyed exactly or not at all.

4 Renders translations whose accuracy is excellent; almost all nuances are conveyed and there are no mistranslations.

4+ Can produce documents that are totally accurate, convey all nuances, and are devoid of mistranslations or omissions.

5 Can produce translations that are an exact reflection of the source document in all aspects, even translating difficult and abstract prose. Can produce work that is totally accurate, with no mistranslations or omissions.

Interpretive information

T-0 NO PROFICIENCY

No ability to translate the language.

T-0+ MEMORIZED PROFICIENCY

Able to translate using only memorized material and expressions, such as numbers, dates, addresses, some street signs and shop designations.

T-1 ELEMENTARY PROFICIENCY (Base Level)

Able to translate very simple documents in printed or typed form at the survival level such as simple messages and simple notes conveying basic instructions.

T-1+ ELEMENTARY PROFICIENCY (Higher Level)

Able to translate simple documents in printed or typed form dealing with survival needs and routine social demands such as simple letters and biographical data.

T-2 LIMITED WORKING PROFICIENCY (Base Level)

Able to produce understandable translations of simple documents pertaining to routine social and business correspondence and areas of professional experience.

T-2+ LIMITED WORKING PROFICIENCY (Higher Level)

Able to translate with some precision most factual, nontechnical prose as well as some documents on concrete topics related to fields in which he or she has an interest or background.

T-3 GENERAL PROFESSIONAL PROFICIENCY
(Base Level)

Able to translate acceptably most formal and informal written exchanges on practical, social and professional topics. Demonstrates an emerging ability to translate diverse subject matter.

T-3+ GENERAL PROFESSIONAL PROFICIENCY
(Higher Level)

Able to translate effectively a variety of documents dealing with diverse subject matter within the scope of personal or professional experience.

T-4 ADVANCED PROFESSIONAL PROFICIENCY
(Base Level)

Able to translate very effectively all forms of documents within the scope of personal and professional experience, can handle other documents adequately.

T-4+ GENERAL PROFESSIONAL PROFICIENCY
(Higher Level)

Approximates a master translator's ability to produce translations that are an exact reflection of the original document.

T-5 (Master Translator Proficiency)

Proficiency equivalent to that of a well-educated master translator. Able to translate even difficult and abstract prose; for example, general technical and legal texts as well as highly colloquial writing.

EXHIBIT A

Paragraph Scoring Grid

(Points)	0 + (0.8)	1 (1.0)	1 + (1.8)	2 (2.0)	2 + (2.8)	3 (3.0)	3 + (3.8)	4 (4.0)	4 + (4.8)	5 (5.0)
Grammer	Almost no attention to basic grammar structures or word order.	Frequent errors in basic grammar structures. Word order generally equivalent to source language.	Many errors in basic grammar structures. Word order very close to source language.	Many errors in high frequency complex grammar structures. Occasional errors in basic structures. Word order closer to source than target language.	Frequent mistakes in high frequency complex grammar structures. Word order begins to approximate that of target language.	Some errors in low frequency complex grammar structures. Some constructions reveal document is a translation.	Occasional errors in low frequency complex grammar structures. Sporadic errors in high-frequency complex grammar structures.	No grammar errors. Word order sometimes typical of source language, but correct.	No grammar errors. Word order normally typical of target language.	No grammar errors. Word order is that of target language. No evidence of translation.
Vocabulary	Mostly nouns (numbers, dates, colors, etc.) and adjectives.	Very simple concrete vocabulary. Nouns, verbs, adjectives.	Everyday vocabulary.	Everyday plus some professional vocabulary.	Active and passive vocabulary, concrete plus some abstract.	Abstract and professional, general and technical, idioms. Uses dictionary well. Uses sayings, proverbs, colloquialisms, but with errors.	Can use most idioms, but lacks comprehensive technical vocabulary. Vocabulary extensive and generally precise.	Equivalent to non-professional native in all general areas. Exact, precise for most occasions. Correct use of idioms, sayings, etc. Aware of register.	Occasional unusual usage of colloquialisms, proverbs, or highly technical vocabulary.	Equivalent to well-educated native. Uses precise vocabulary to convey nuances. Can use literary, technical, classical terms and cultural references.
Spelling	Less than 50% correct.	50-65%	66-70%	71-75%	76-79%	80-85%	86-89% correct	90-95%	96-98%	99% correct
Punctuation	Less than 50% correct.	50-65%	66-70%	71-75%	76-79%	80-85%	86-89% correct	90-95%	96-98%	99% correct
Accuracy	Many mistranslations and omissions.	Many mistranslations and omissions.	Frequent mistranslations and omissions.	Quite a few mistranslations. Many omissions. Some additions.	Some mistranslations and omissions. Limited ability to convey nuances.	Occasional mistranslations and omissions. Some nuances not conveyed.	Infrequent mistranslations. Occasional omissions or missed nuances.	No mistranslations. Omissions very rare. Almost all nuances conveyed.	No mistranslations or omissions. All nuances conveyed.	No mistranslations or omissions. All nuances conveyed.
Style	Distorted; not identifiable.	Uneven. Varies widely from original.	Uneven. Does not reflect original.	Original frequently distorted. Uneven.	Occasionally distorted and uneven.	Occasionally differs from original.	Varies slightly from original.	Equivalent to original.	Exactly reflects original. Adequate for almost all subject matter. Smooth, fluid.	Exactly reflects original. Adequate for all subject matter. Smooth, fluid.
Tone	Distorted; not identifiable.	Uneven. Varies widely from original.	Uneven. Does not reflect original.	Original frequently distorted. Uneven.	Occasionally distorted and uneven.	Occasionally differs from original.	Varies slightly from original.	Equivalent to original.	Exact reflection of original.	Exact reflection, including humor.

Exhibit A

EXHIBIT B

QUESTIONNAIRE ON TRANSLATION SKILL LEVELS

Please read the attached information on translation skill levels. We ask that you examine the criteria, descriptions, and scoring grid in light of your experience with translation. Your comments on this material will help us to develop an accurate test of translation ability. If you require more space than is provided after each question, please continue your responses on the back.

Section A. Criteria

1. What relationship do you see between ILR reading/writing level and translation skill level? Do you agree with the assessment of the relationship described in the criteria?

2. Do you agree with the description of a "perfect" translation? Why or why not?

3. Are there variables other than those presented that you would consider in evaluating translation ability? Do you consider any of the variables presented to be unimportant?

Section B. Translation Level Descriptions

Please read through each skill level description and note any comments regarding a particular description in your responses to the questions below. Be sure to indicate the skill level description and the line within that description that your comment applies to.

1. Do you think any of the characteristics we have included in Level 0-5 is inappropriate to that level? If so, which?

2. Where would you add other characteristics?

3. Would you delete any characteristics from the descriptions?

4. Are there unclear areas in any of the descriptions? ________

5. Do you agree with the description of a Master Translator?

6. **What would you add to, change, or delete from this description (T-5)?**

Section C. Scoring Grid

The attached grid is designed to aid scorers in making a decision about the appropriate skill level description to assign. Please comment on the grid.

1. Would you find this grid helpful in evaluating a translation test? ________

2. Where would you make changes to the grid? ____________________

____

____

____

3. What would you add to the grid? ____________________

____

____

____

4. Do you agree with the percentages listed for spelling and punctuation accuracy? If not, what percentages would you substitute? ____________________

____

____

____

We would welcome any additional comments you might have. Please use the rest of this page or an additional sheet to comment on any aspect of this material. Thank you for your valuable assistance in developing criteria for rating tests of translation ability.

Sincerely,

Charles Stansfield
Marijke Walker

APPENDIX J

BACKGROUND PROFICIENCY QUESTIONNAIRE
GIVEN BEFORE TRIALING

Name: ______________________________

Date: ______________

Test: ______________________________

Thank you very much for agreeing to take part in the trialing of the Spanish into English Verbatim Translation Exams. Your comments about these exams are very important to us. We would like you to fill out these forms after you have completed each version of the exam. Please be as clear and frank as possible.

The exact time for completing each section has not yet been established but we would like you to work as quickly and accurately as you can (as if it were a timed exam). Please record the time needed to complete each section on these forms. This will enable us to establish the completion times for future examinees.

You are not permitted to use a dictionary on any part of this exam except for the last section which is entitled "Production Section III." You are also not permitted to receive or give any assistance regarding these exams. Your cooperation in these matters is greatly appreciated.

How do you rate your overall Spanish ability?

How do you rate your overall English ability?

APPENDIX K

EXAM FEEDBACK QUESTIONNAIRE
MULTIPLE CHOICE AND PRODUCTION SECTIONS
(TRIALING VERSION)

Multiple Choice Section I Completion time:_____hrs. _____minutes

1) How could the directions be made clearer?

2) How should questions be modified, if any, so that they are less misleading/confusing?

3) Which questions, if any, do you feel should be deleted?

4) Which questions, if any, do you feel should be added?

5) What unintended errors, if any, did you find in this section?

6) Did this section adequately test your knowledge of English?

7) Did this section adequately test your knowledge of Spanish?

8) Were any major points not tested that you feel should have been?

9) Did you feel that this section was: too long / too short / just right?

10) Any additional comments? (Continue on the back, if necessary!!)

Multiple Choice Section II Completion time:____hrs. ____minutes

1) How could the directions be made clearer?

2) How should questions be modified, if any, so that they are less misleading/confusing?

3) Which questions, if any, do you feel should be deleted?

4) Which questions, if any, do you feel should be added?

5) What unintended errors, if any, did you find in this section?

6) Did this section adequately test your knowledge of English?

7) Did this section adequately test your knowledge of Spanish?

8) Were any major points not tested that you feel should have been?

9) Did you feel that this section was: too long / too short / just right?

10) Any additional comments? (Continue on the back, if necessary!!)

Production Section I Completion time:____hrs. ____minutes

1) How could the directions be made clearer?

2) How should questions be modified, if any, so that they are less misleading/confusing?

3) Which questions, if any, do you feel should be deleted?

4) Which questions, if any, do you feel should be added?

5) What unintended errors, if any, did you find in this section?

6) Did this section adequately test your knowledge of English?

7) Did this section adequately test your knowledge of Spanish?

8) Were any major points not tested that you feel should have been?

9) Did you feel that this section was: too long / too short / just right?

10) Any additional comments? (Continue on the back, if necessary!!)

Production Section II Completion time:_____hrs. _____minutes

1) How could the directions be made clearer?

2) How should questions be modified, if any, so that they are less misleading/confusing?

3) Which questions, if any, do you feel should be deleted?

4) Which questions, if any, do you feel should be added?

5) What unintended errors, if any, did you find in this section?

6) Did this section adequately test your knowledge of English?

7) Did this section adeuately test your knowledge of Spanish?

8) Were any major points not tested that you feel should have been?

9) Did you feel that this section was: too long / too short / just right?

10) Any additional comments? (Continue on the back, if necessary!!)

APPENDIX L

SEVTE EXAM FEEDBACK QUESTIONNAIRE
(VALIDATION STUDY)

SPANISH INTO ENGLISH VERBATIM EXAM QUESTIONNAIRE

We would very much appreciate your answers to the following brief questions concerning the verbatim translation exams you have just taken:

1. *Was the length of time given for completing the multiple choice sections about right?*

   ( ) Too short
   ( ) About right
   ( ) Too long

2. *Was the length of time given for completing the production sections about right?*

   ( ) Too short
   ( ) About right
   ( ) Too long

Please indicate to what extent you agree or disagree with the following statements:

3. *The directions were clear.*

   ( ) Agree *( ) Disagree*

4. *The material in the exams was representative of the types of written documents I might encounter in my work.*

   ( ) Strongly agree *( ) Agree* *( ) Disagree* *( ) Strongly disagree*

5. *There was sufficient opportunity for me to demonstrate my ability to translate from Spanish into English.*

   ( ) Strongly agree *( ) Agree* *( ) Disagree* *( ) Strongly disagree*

Thank you for your cooperation.

APPENDIX M

PILOT QUESTIONNAIRE AND RESULTS
ON
LANGUAGE BACKGROUND AND PROFICIENCY

Thank you for agreeing to assist us in evaluating these tests. We request that you complete the following information to aid in our analysis.

Name: ______________________________

Profession:

___ Student
Course of Study:
___ Bachelor's in Spanish
___ Master's in Spanish
___ Translation Certificate Program
___ Other (Please specify)

___ Translator
___ Teacher
___ Other (please specify)

Native Language:

___ English
___ Spanish
___ Other (please specify)

How would you rate your ability to write in English?

___ Excellent
___ Very good
___ Good
___ Fair
___ Poor

How would you rate your ability to speak in English?

___ Excellent
___ Very good
___ Good
___ Fair
___ Poor

How would you rate your ability to write in Spanish?

___ Excellent
___ Very good
___ Good
___ Fair
___ Poor

How would you rate your ability to speak in Spanish?

___ Excellent
___ Very good
___ Good
___ Fair
___ Poor

QUESTIONNAIRE RESULTS

UNDERGRADUATES

Total Respondents: 45

All data self-reported

Native Language:

English: 38

Spanish: 0

Bilingual
Eng-Span: 1

Other: 6

English Writing Ability:

Excellent: 22
Very good: 16
Good: 6
Fair: 1
Poor: 0

English Speaking Ability:

Excellent: 29
Very good: 15
Good: 0
Fair: 1
Poor: 0

Spanish Writing Ability:

Excellent: 1
Very good: 9
Good: 20
Fair: 12
Poor: 3

Spanish Speaking Ability:

Excellent: 2
Very good: 6
Good: 16
Fair: 18
Poor: 3

GRADUATE STUDENTS

Total Respondents: 10

All data self-reported

Native Language:

English: 3

Spanish: 6

Bilingual
Eng-Span: 0

Other: 1

English Writing Ability:

Excellent: 1
Very good: 6
Good: 3
Fair: 0
Poor: 0

English Speaking Ability:

Excellent: 3
Very good: 4
Good: 3
Fair: 0
Poor: 0

Spanish Writing Ability:

Excellent:	4
Very good:	3
Good:	1
Fair:	2
Poor:	0

Spanish Speaking Ability:

Excellent:	5
Very good:	2
Good:	2
Fair:	1
Poor:	0

APPENDIX N

SELF-ASSESSMENT QUESTIONNAIRE

AND

SUMMARY REPORT ON SELF-ASSESSMENT

NAME ____________________ FIELD OFFICE ________________

SELF-ASSESSMENT OF TRANSLATION ABILITY

The purpose of this questionnaire is to learn your candid evaluation of your ability to translate written documents from SPANISH INTO ENGLISH. It is of the utmost importance that you provide an honest evaluation of your present abilities so that the effectiveness of the translation exams may be accurately and fully assessed. Please be assured that your responses will be kept confidential by the test development contractor and will in no way affect your standing or possibility of advancement within the Bureau.

Instructions: Please estimate your ability to translate the following types of documents using the scale provided below:

Limited	The translated document contains many mistranslations and omissions, and frequent errors in grammar. The translation is extremely literal (i.e. word for word) and may be difficult to understand.
Functional	The translation is fairly accurate with no substantive omissions; however, it may contain some mistranslations and grammar errors. The translation is literal but generally understandable.
Competent	The accuracy of the translated document is good, with occasional minor mistranslations and omissions. There is no pattern of grammar errors. Most idiomatic expressions are used appropriately; however, the phrasing may reveal the document to be a translation.
Superior	The accuracy of the translation is excellent, with most nuances conveyed. Grammar errors are rare. The phrasing is entirely natural and the document does not appear to be a translation.

Please evaluate candidly your ability to translate each of the following types of documents from Spanish into English by circling the appropriate label. If you have never translated a particular type of document, please mark N/A ("not applicable").

1. Newspaper articles	Limited	Functional	Competent	Superior	N/A
2. Newspaper editorials	Limited	Functional	Competent	Superior	N/A
3. Depositions	Limited	Functional	Competent	Superior	N/A
4. Police reports	Limited	Functional	Competent	Superior	N/A
5. Correspondence	Limited	Functional	Competent	Superior	N/A
6. Legal documents	Limited	Functional	Competent	Superior	N/A
7. Letters rogatory	Limited	Functional	Competent	Superior	N/A
8. Case histories	Limited	Functional	Competent	Superior	N/A
9. FCI status/evaluation reports	Limited	Functional	Competent	Superior	N/A
10. Scientific/technical articles	Limited	Functional	Competent	Superior	N/A
11. Foreign diplomatic reports	Limited	Functional	Competent	Superior	N/A
12. Training manuals	Limited	Functional	Competent	Superior	N/A
13. ____________________ (Please specify)	Limited	Functional	Competent	Superior	N/A

SUMMARY REPORT ON SELF-ASSESSMENT: SPANISH TO ENGLISH

The following section is an analysis of the results of the Spanish-to-English Self-Assessment Questionnaire that was completed by FBI personnel participating in the validation study.

This section specifies:

1. the document types which the participants checked most frequently;
2. the average rating for each document type;
3. the per cent of the total respondents who gave a response for each document type;
4. the document types which correlated most significantly with the FBI translation skill level descriptions.

AVERAGE RATING OF EACH DOCUMENT TYPE

The questionnaire required the employee to rate his or her ability to translate each document type on a four point scale. The options on the scale were: 4, superior; 3, competent; 2, functional; and 1, limited. The documents listed below were included. In addition, there were 43 respondents to the Spanish-to-English self-assessment questionnaire. The table below gives the percent who responded to each document type, and the average rating, ranked in descending order.

DOCTYPE	% RESPNDING	AVERAGE SELF-RATING
SECORRES(correspondence)	98	3.11
SENEWSAR(newspaper articles)	86	3.02
SEDEPOS(depositions)	58	3.00
SENEWSED(news editorials)	81	2.94
SEPOLRPT(police reports)	77	2.93
SELETROG(letters rogatory)	58	2.88
SETRNG(training manuals)	49	2.85
SECASHST(cash statements)	56	2.83
SELEGAL(legal documents)	70	2.70
SEDIPL(foreign diplomatic	47	2.70
SEFCI(FCI reports)	49	2.61
SETECH(technical articles)	53	2.43

The self-rating most frequently chosen was COMPETENT, except in the case of technical documents, where an equal number of respondents chose FUNCTIONAL as their self-rating. News articles, editorials and correspondence were the document types most frequently chosen.

CORRELATIONS WITH OVERALL SCORES

The table below presents the correlations of each document type with the overall scores for Expression and Accuracy. The number of paired scores is listed in parentheses below each correlation:

DOCTYPE	EXPF1	EXPF2	ACCF1	ACCF2
SENEWSAR	0.30 (37)	0.22 (36)	0.50* (37)	0.46* (36)
SENEWSED	0.27 (35)	0.22 (34)	0.57* (35)	0.51* (34)
SEDEPOS	0.57* (25)	0.40 (24)	0.73* (25)	0.72* (24)
SEPOLRPT	0.43* (33)	0.30 (32)	0.56* (33)	0.56* (32)
SECORRES	0.41* (42)	0.27 (41)	0.59* (42)	0.64* (41)
SELEGAL	0.43* (30)	0.20 (29)	0.55* (30)	0.50* (29)
SELETROG	0.51* (25)	0.39* (25)	0.54* (25)	0.62* (25)
SECASHST	0.39 (24)	0.21 (24)	0.52* (24)	0.50* (24)
SEFCI	0.53* (21)	0.24 (21)	0.65* (21)	0.57* (21)
SETECH	0.54* (23)	0.23 (22)	0.50* (23)	0.42* (22)
SEDIPL	0.64* (20)	0.38 (19)	0.73* (20)	0.74* (19)
SETRNG	0.48* (21)	0.24 (21)	0.53* (21)	0.66* (21)

*p<.05

On Form 1, the documents showing the highest correlations for Expression were, in descending order: foreign diplomatic reports, depositions, technical manuals, letters rogatory and FCI reports. On Form 2, only letters rogatory showed any significant correlation, which was less than 0.50. By comparison, Accuracy total correlations were both higher and more frequent.

On Form 1, the documents showing the highest correlation for Accuracy were, in descending order: foreign diplomatic reports and depositions (with the same correlation of 0.73); FCI reports, correspondence, news editorials, and police reports. On Form 2, these documents were foreign diplomatic reports, depositions, training manuals, correspondence, letters rogatory, FCI reports, and police reports.

The magnitude and the order of the correlations for each type of translation task was almost identical across the two forms, suggesting that the two forms are consistent in their criterion-related validity.

APPENDIX O

CONVERSION TABLES: RAW SCORE TO TSL SCORE

EXPRESSION AND ACCURACY

Conversion Table

Expression Raw Score	TSL Score
1	*
2	*
3	*
4	*
5	*
6	*
7	*
8	*
9	*
10	*
11	*
12	*
13	*
14	*
15	*
16	0.4
17	0.5
18	0.5
19	0.5
20	0.6
21	0.6
22	0.7
23	0.7
24	0.8
25	0.8
26	0.9
27	0.9
28	0.9
29	1.0
30	1.0
31	1.1
32	1.1
33	1.2
34	1.2
35	1.3
36	1.3
37	1.3
38	1.4
39	1.4
40	1.5
41	1.5
42	1.6
43	1.6
44	1.7
45	1.7
46	1.7
47	1.8
48	1.8

* 1-15 = chance scores

Expression Raw Score	TSL Score
49	1.9
50	1.9
51	2.0
52	2.0
53	2.1
54	2.1
55	2.1
56	2.2
57	2.2
58	2.3
59	2.3
60	2.4
61	2.4
62	2.5
63	2.5
64	2.5
65	2.6
66	2.6
67	2.7
68	2.7
69	2.8
70	2.8
71	2.9
72	2.9
73	2.9
74	3.0
75	3.0
76	3.1
77	3.1
78	3.2
79	3.2
80	3.3
81	3.3
82	3.3
83	3.4
84	3.4
85	3.5
86	3.5
87	3.6
88	3.6
89	3.7
90	3.7
91	3.7
92	3.8
93	3.8
94	3.9
95	3.9
96	4.0
97	4.0
98	4.1
99	4.1
100	4.2

Expression Raw Score	TSL Score
101	4.2
102	4.2
103	4.3
104	4.3
105	4.4

Form 1 - SEVTE

Conversion Tables

Accuracy Raw Score	TSL Score
1	0.6
2	0.7
3	0.7
4	0.8
5	0.8
6	0.9
7	0.9
8	0.9
9	1.0
10	1.0
11	1.1
12	1.1
13	1.2
14	1.2
15	1.3
16	1.3
17	1.4
18	1.4
19	1.4
20	1.5
21	1.5
22	1.6
23	1.6
24	1.7
25	1.7
26	1.8
27	1.8
28	1.9
29	1.9
30	1.9
31	2.0
32	2.0
33	2.1
34	2.1
35	2.2
36	2.2
37	2.3
38	2.3
39	2.4
40	2.4
41	2.4
42	2.5
43	2.5
44	2.6
45	2.6
46	2.7
47	2.7
48	2.8

Accuracy Raw Score	TSL Score
49	2.8
50	2.9
51	2.9
52	2.9
53	3.0
54	3.0
55	3.1
56	3.1
57	3.2
58	3.2
59	3.3
60	3.3
61	3.3
62	3.4
63	3.4
64	3.5
65	3.5
66	3.6
67	3.6
68	3.7
69	3.7
70	3.8
71	3.8
72	3.8
73	3.9
74	3.9
75	4.0
76	4.0
77	4.1
78	4.1
79	4.2
80	4.2

Conversion Table

Expression Raw Score	TSL Score
1	*
2	*
3	*
4	*
5	*
6	*
7	*
8	*
9	*
10	*
11	*
12	*
13	*
14	*
15	*
16	0.6
17	0.6
18	0.7
19	0.7
20	0.8
21	0.8
22	0.8
23	0.9
24	0.9
25	1.0
26	1.0
27	1.0
28	1.1
29	1.2
30	1.2
31	1.3
32	1.3
33	1.3
34	1.4
35	1.4
36	1.5
37	1.5
38	1.6
39	1.6
40	1.7
41	1.7
42	1.8
43	1.8
44	1.8
45	1.9
46	1.9

* 1-15 = chance scores

Expression Raw Score	TSL Score
47	2.0
48	2.0
49	2.0
50	2.1
51	2.2
52	2.2
53	2.3
54	2.3
55	2.4
56	2.4
57	2.4
58	2.5
59	2.5
60	2.6
61	2.6
62	2.7
63	2.7
64	2.8
65	2.8
66	2.9
67	2.9
68	3.0
69	3.0
70	3.0
71	3.1
72	3.1
73	3.2
74	3.2
75	3.3
76	3.3
77	3.4
78	3.4
79	3.4
80	3.5
81	3.5
82	3.6
83	3.6
84	3.7
85	3.7
86	3.8
87	3.8
88	3.9
89	3.9
90	4.0
91	4.0
92	4.1
93	4.1
94	4.1
95	4.2
96	4.2
97	4.3

Expression Raw Score	TSL Score
98	4.3
99	4.4
100	4.4
101	4.4
102	4.5
103	4.5
104	4.6
105	4.6

Conversion Tables

Accuracy Raw Score	TSL Score
1	0.2
2	0.3
3	0.3
4	0.4
5	0.4
6	0.5
7	0.5
8	0.6
9	0.6
10	0.7
11	0.8
12	0.8
13	0.9
14	0.9
15	1.0
16	1.0
17	1.1
18	1.1
19	1.2
20	1.2
21	1.3
22	1.4
23	1.4
24	1.5
25	1.5
26	1.6
27	1.6
28	1.7
29	1.7
30	1.8
31	1.9
32	1.9
33	2.0
34	2.0
35	2.1
36	2.1
37	2.2
38	2.2
39	2.3
40	2.3
41	2.4
42	2.5
43	2.5
44	2.6
45	2.6

Accuracy Raw Score	TSL Score
46	2.7
47	2.7
48	2.8
49	2.8
50	2.9
51	3.0
52	3.0
53	3.1
54	3.1
55	3.2
56	3.2
57	3.3
58	3.3
59	3.4
60	3.4
61	3.5
62	3.5
63	3.6
64	3.7
65	3.7
66	3.8
67	3.8
68	3.9
69	3.9
70	4.0
71	4.1
72	4.1
73	4.2
74	4.2
75	4.3
76	4.3
77	4.4
78	4.4
79	4.5
80	4.5

APPENDIX P

MEMORANDUM ON TOTAL SCORE CONVERSION
TO
FBI/CAL EQUIVALENCY RATING

Memo
To: Marijke Walker
From: Charles Stansfield
Date: May 15, 1990
Subject: Total score conversion to ILR equivalency rating

As I indicated to you on the phone, we have encountered a problem in converting the total score on the test to an ILR-like Translation Rating. Each examinee took two forms of the test and each examinee was given an overall ILR-like rating by each of two raters based on the examinee's performance on each test. The raters assigned ratings for Accuracy and Expression. Thus, each examinee received four estimates of his ILR level (estimates per form) for accuracy and four estimates of his ILR level for expression.

We averaged the four estimates of ILR rating to come up with an overall Translation rating. We then correlated the test scores with the Translation rating. The high correlation (an average of .90) allowed us to use the resulting regression equation to predict Translation rating from the total score on the test. Thus, we were able to construct a score conversion table for all points on the test scale which would produce an estimated Translation skill level.

One of the problems with such conversion tables is a phenomenon known as the "regression effect" (different meaning from the use of regression above). The regression effect means that examinee's whose first score is far from the mean will be predicted to be closer to the mean on the second score. Thus, most examinees whose score on our test is at the top of the distribution will be predicted to have a lower ILR score than they received from the raters. Similarly, most examinees whose score on our test was at the bottom of the distribution were predicted to have a higher ILR score than they received from the raters.

Attached is a copy of the scatterplot for 42 FBI examinees. The ILR expression rating is on the vertical axis, while the total expression score on our test (ESVTE) is on the horizontal axis. We have drawn in the regression line with a pencil. This is the straight line that best fits the distribution. For any other line, if you calculated the deviations produced by comparing obtained scores with the predicted scores, the sum of the deviations from the regression line would be greater.

On this scatterplot each A represents one examinee. Each B represents two examinees. As indicated in the note at the bottom, 14 examinees' scores are not on the scatterplot because their scores and the regression line coincided. Thus, for these examinees, the conversion table worked perfectly. The asterisks are the computer's representation of the regression line. In this scatterplot you will see some tendency for the deviations between the actual and predicted score to be quite small near the center of the distribution, and larger at the ends. You will also see some tendency for examinees who scored above 80 on the ESVTE to have a predicted score that is lower than their obtained score. Similarly, for examinees who scored below 40, the predicted score is usually higher than the obtained score. Thus, more of the obtained scores for these people are below the regression line than above it.

One effect of the regression effect is to lower the range of ability measured by the test. That is, the highest ability examinee on this test obtained a rating of 4.5 but the conversion table predicts his predicted skill level to be 3.8. This person was probably one of the three professional translators who took the test.

One option we have, which would reduce the regression effect described in paragraph three above is to tilt the regression line to the left by transforming the scores so that the maximum ILR score level is higher, 4.5 for example. However, we have no basis other than intuition for doing this. That is, the sample did not contain people whom we knew beforehand were at the 4.5 level or higher. While this seems reasonable, in that it reduces the regression effect, it also increases slightly the amount of error in the predicted ILR scores all along the continuum. Thus, it seems unwise.

Another option is to have several people take the test whom we know to be level 4+ and 5 translators, and enter their results into the equation. This would have to be done later, however. So, that's our dilemma. As it stands, no one in the sample would earn a predicted ILR rating above 3+, and because of the lack of high ability examinees in the sample, it is not possible to earn a rating higher than 4.2 on the test, even though we believe it to be sensitive to differences in ability in the 4-5 range. Further evidence that the test could discriminate in that range is found in the fact that the highest raw Expression score on the test was 98 on the ESVTE and 96 on the SEVTE, while the maximum possible total score was 105. Similarly, for Accuracy, the highest raw score was 71 on the SEVTE and 75 for the ESVTE, while the maximum possible total score was 80. Thus, the difficulty level of the test exceeds the ability level of any examinee in the sample.

As a future project, we should think about how we can identify at least 10 high level translators and then administer the tests to them. We would then be able to revise the score conversion table so that the ILR ratings for high ability candidates are more accurate than at present, and so that the test will measure ability up to a higher level than at present.

For the moment, it may be best to leave the conversion table as is. However, if this conversion table is used, test score users should be aware that it may underpredict the true levels of examinees whose predicted ILR rating is 3.5 or above. This information should be incorporated in any test manual that you prepare.

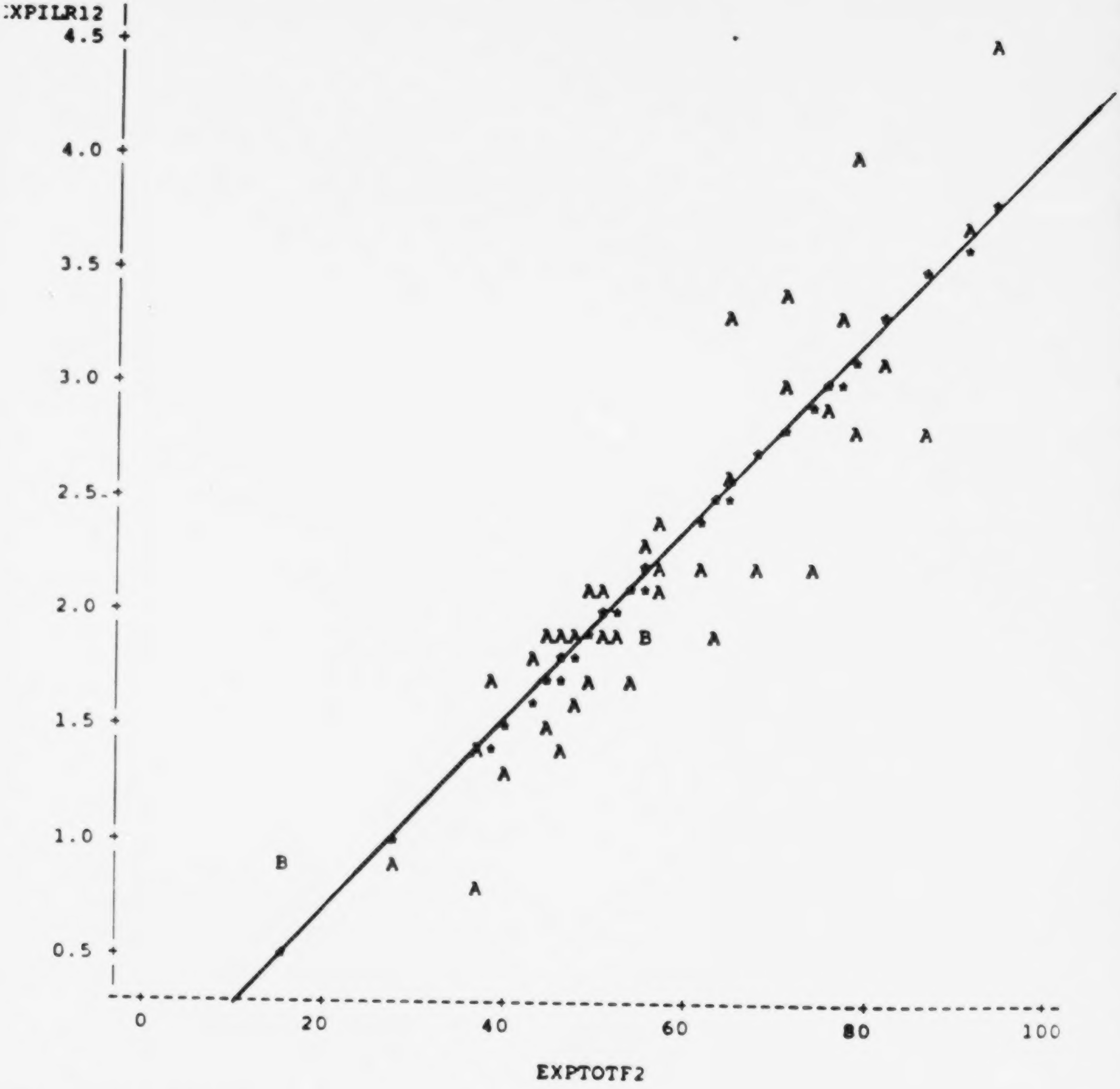
In general, I find this disappointing. We tried to make the test hard enough to measure ability as high as level 5. However, because 5's did not show up in the sample, the test appears to fail to measure at such a high level.

On a more positive note, I should say that the test seems to predict the average Translation skill level rating assigned by our raters very accurately between the 1.8 and 3.5 range, which is the range in which most FBI personnel scored.

I should mention one more concern. All of the 17 FBI employees on whom we had Translation level ratings on the FBI's current translation test received a lower Translation rating on our test than on the FBI test. The average difference was about half a full level, with differences typically being larger for examinees whose FBI test score was 3.8 or above, and being smaller for examinees whose FBI test score was 2.8 or below. Thus, either a.) the FBI's current test is too generous, or b.) our raters are too severe, or c.) the time constraints on our test do not permit the examinees to revise their translations and demonstrate their true ability, or d.) the examinees were not motivated to give their best performance when they took our test, or e.) the examinees' true Translation ability declined subsequent to taking the FBI test. Do you have any thoughts about a.) or e.) above?

```
                 ENSP Form 2: EXPILR12 Predicted from exptotf2                    56
                                            13:57 Tuesday, May 15, 1990

     Plot of EXPILR12*EXPTOTF2.  Legend: A = 1 obs, B = 2 obs, etc.
     Plot of PRED*EXPTOTF2.      Symbol used is '*'.
```



```
TE: 14 obs hidden.
```

APPENDIX Q

SURVEY
OF
FBI TRANSLATION NEEDS

Dear Language Specialist,

The Language Services Unit has contracted with the Center for Applied Linguistics (CAL) to develop a new translation test, Spanish into English and English into Spanish. We would like to develop a new test which tests more closely for the actual linguistic tasks carried out by Language Specialists. Therefore, we would really appreciate your input. We kindly ask you to fill out the attached questionnaire; feel free to add any comments you think are pertinent. Please note that "% OF YOUR TIME" refers to the percentage of time that is devoted to the listed tasks when you are working with the Spanish language, and NOT to the percentage of time that is DEVOTED TO THE TASKS OUT OF YOUR WORKDAY. This becomes a pertinent difference especially for those of you who work with a number of languages. To illustrate this point, a certain language specialist may devote roughly half of his time in his Spanish-language work to interpretation assignments, but his work with the Spanish language itself might constitute only a fraction of his entire workday.

If an item does not apply to you, put 0 % in the appropriate column. As concerns the other (please specify) listing, please note that we are interested only in tasks that are performed on a regular basis. There is no need for you to list any assignment that was performed once or that is performed only rarely.

Please return the completed questionnaires to me as soon as possible (Bureau mail); an addressed envelope has been attached for this purpose.

Thank you so much for your help.

Marijke Walker
Testing Program Manager
Language Services Unit
FBIHQ, Room 3505

Phone HQ x4160

QUESTIONNAIRE TO DETERMINE THE FBI'S TRANSLATION NEEDS

FROM ENGLISH TO SPANISH

I. ORAL TASKS

% OF YOUR TIME

Interpretation Assignments ____________

Check as many as are applicable
____ unannounced visitors
____ tours
____ conferences
____ other (please specify)

Oral Proficiency Test (Spanish) ____________

II. TASKS INVOLVING WRITTEN MATERIAL

	% OF YOUR TIME TRANSLATING	% OF YOUR TIME SUMMARIZING
Legal Documents	____________	____________

Check as many as are applicable
____ letters rogatory
____ extradition requests
____ laws/violations/legal rights
____ wanted posters
____ other (please specify)

Booklets Manuals	____________	____________

Check as many as are applicable
____ science technology
____ tours
____ training
____ other (please specify)

Forms	____________	____________

Check as many as are applicable
____ Bureau forms
____ DOJ forms
____ other (please specify)

Other (please specify)	____________	____________

% OF YOUR TIME SPENT IN TRANSLATING　　% OF YOUR TIME SPENT IN SUMMARIZING

Recorded Conversations:

TELEPHONE ____________ ____________

CHECK AS MANY AS ARE APPLICABLE:

____ politics
____ business/finance
____ economics
____ general theft/white collar crime
____ organized crime
____ narcotics trafficking
____ domestic/international terrorism
____ foreign counterintelligence
____ science/technology
____ military
____ legal
____ theft
____ gambling
____ counterfeiting
____ kidnapping
____ procedures/appointments
____ payments/purchases
____ explanations
____ other (please specify)

BODY RECORDER ____________ ____________

CHECK AS MANY AS ARE APPLICABLE:

____ politics
____ business/finance
____ economics
____ general theft/white collar crime
____ organized crime
____ narcotics trafficking
____ domestic/international terrorism
____ foreign counterintelligence
____ science/technology
____ military
____ legal
____ theft
____ gambling
____ counterfeiting
____ kidnapping
____ procedures/appointments
____ payments/purchases
____ explanations
____ other (please specify)

Other (please specify): ____________ ____________

	% OF YOUR TIME SPENT IN TRANSLATING	% OF YOUR TIME SPENT IN SUMMARIZING
Medical Reports	________	________

CHECK AS MANY AS ARE APPLICABLE:

____ autopsies
____ other (please specify)

Patents	________	________
Other (please specify):	________	________

IV. TASKS INVOLVING LISTENING

	% OF YOUR TIME SPENT IN TRANSLATING	% OF YOUR TIME SPENT IN SUMMARIZING
Broadcasts:	____ ____	________

CHECK AS MANY AS ARE APPLICABLE:

____ politics
____ business/finance
____ economics
____ general theft/white collar crime
____ organized crime
____ narcotics trafficking
____ domestic/international terrorism
____ foreign counterintelligence
____ science/technology
____ military
____ legal
____ other (please specify)

	% OF YOUR TIME SPENT IN TRANSLATING	% OF YOUR TIME SPENT IN SUMMARIZING
Domestic/International Terrorism	____________	____________

CHECK AS MANY AS ARE APPLICABLE:

____ status and evaluation reports
____ case histories
____ police records
____ court records
____ travel documents
____ other (please specify)

Foreign Counterintelligence	____________	____________

CHECK AS MANY AS ARE APPLICABLE:

____ status and evaluation reports
____ material on
intelligence communication methods
____ case histories
____ notices of assignment of diplomats
____ other (please specify)

Treaty Requests/Letters Rogatory	____________	____________
Scientific/Technical	____________	____________

CHECK AS MANY AS ARE APPLICABLE:

____ chemistry
____ biology
____ fingerprinting/DNA typing
____ computer technology
____ explosive and incendiary devices
____ weapons
____ automobiles and other vehicles
____ other (please specify)

	% OF YOUR TIME SPENT IN TRANSLATING	% OF YOUR TIME SPENT IN SUMMARIZING
Letters to the Director and other FBI officials:	____________	____________
Teletypes: (TRANSLATION ONLY)	____________	____________

Legal/Technical:

General Theft/White Collar Crime	____________	____________

CHECK AS MANY AS ARE APPLICABLE:

____ bank records
____ police reports
____ court records
____ other (please specify)

Organized Crime	____________	____________

CHECK AS MANY AS ARE APPLICABLE:

____ status and evaluation reports
____ bank records
____ police reports
____ court records
____ other (please specify)

Narcotics Trafficking	____________	____________

CHECK AS MANY AS ARE APPLICABLE:

____ status and evaluation reports
____ bank records
____ police reports
____ court records
____ other (please specify)

QUESTIONNAIRE TO DETERMINE THE FBI'S TRANSLATION NEEDS

ROM SPANISH INTO ENGLISH

% OF YOUR TIME

ORAL TASKS

Interpretation Assignments: ______________

CK AS MANY AS ARE APPLICABLE:

___ unannounced visitors
___ tours
___ conferences
___ other (please specify)

Oral Proficiency Examinations: ______________
(ENGLISH ONLY)

% OF YOUR TIME

. GRADING OF FOREIGN LANGUAGE EXAMINATIONS ______________

I. TASKS INVOLVING WRITTEN MATERIAL

	% OF YOUR TIME SPENT IN TRANSLATING	% OF YOUR TIME SPENT IN SUMMARIZING
wspapers/Magazines:	______________	______________

CK AS MANY AS ARE APPLICABLE:

___ news items
___ editorials

cles on

___ politics
___ business/finance
___ economics
___ general theft/white collar crime
___ organized crime
___ narcotics trafficking
___ domestic/international terrorism
___ foreign counterintelligence
___ science/technology
___ military
___ legal
___ other (please specify)

QUESTIONNAIRE RESULTS

TOTAL NUMBER OF RESPONDENTS:	28

AVERAGE TIME SPENT

(Averages were calculated based on number of respondents to each question; 0% answers were not factored in unless all answers were 0)

ORAL TASKS

Interpretation Assignments

Number of respondents:	19/28
Average % of time spent	4.8%

The most frequent category checked by respondents was "unannounced visitors." Under "other," respondents listed tasks such as interviewing suspects, handling complaints, and debriefing informants, witnesses and subjects.

Oral Proficiency Examinations

Number of respondents:	1/28
Average % of time spent	1.0%

GRADING OF FOREIGN LANGUAGE EXAMINATIONS

Number of respondents:	1/28
Average % of time spent	70.0%

TASKS INVOLVING WRITTEN MATERIAL

Newspapers/Magazines

% of time spent translating	% of time spent summarizing
23.3%	21.0%

Number of respondents	Number of respondents
12/28	5/28

The categories most chosen by respondents were politics, narcotics, terrorism, foreign counterintelligence, legal, theft, and organized crime. The other categories were seldom chosen.

Letters to the Director and other FBI officials

% of time spent translating	% of time spent summarizing
1.8%	2%

Number of respondents	Number of respondents
4/28	1/28

Teletypes

% of time spent translating	% of time spent summarizing
1.0%	0%

Number of respondents	Number of respondents
1/28	0/28

Legal/Technical

General Theft/White Collar Crime

% of time spent translating	% of time spent summarizing
9.75%	11%

Number of respondents	Number of respondents
12/28	2/28

All categories were chosen by respondents. Under "other," translation of letters was indicated, as well as translation of affidavits and signed statements. These "other" items were repeated throughout this section.

Organized Crime

% of time spent translating	% of time spent summarizing
8.1%	5%
Number of respondents	Number of respondents
9/28	1/28

The category most frequently chosen was "police reports."

Narcotics Trafficking

% of time spent translating	% of time spent summarizing
17.1%	37.5%
Number of respondents	Number of respondents
15/28	4/28

The category most frequently chosen was "court records." Under "other," translation of letters and ledger (log) notes was indicated, as were T-III and T-IV translations.

Domestic/International Terrorism

% of time spent translating	% of time spent summarizing
13.2%	25.5%
Number of respondents	Number of respondents
10/28	2/28

The most frequent responses were "case histories" and "court records." Among "other" responses was translation of communiqués.

Foreign Counterintelligence

% of time spent translating	% of time spent summarizing
18.6%	24.4%
Number of respondents	Number of respondents
18/28	7/28

The category most frequently chosen was "status and evaluation reports." Under "other," categories listed include political and military intelligence and defectors' reports.

Treaty Requests/Letters Rogatory

% of time spent translating	% of time spent summarizing
.75%	0
Number of respondents	Number of respondents
2/28	0/28

Scientific/Technical

% of time spent translating	% of time spent summarizing
12%	0
Number of respondents	Number of respondents
6/28	0

The categories most frequently chosen were explosive and incendiary devices, weapons, and autombiles and other vehicles. Fingerprinting/DNA typing and computer technology were seldom chosen.

Medical Reports

% of time spent translating	% of time spent summarizing
3.9%	0

Number of respondents	Number of respondents
8/28	0

"Other" responses include medical reports to be used as evidence, progress reports, and hospital reports.

Patents

Number of respondents	Number of respondents
0/28	0

Other (Respondent listed police reports and ownership/sale documents).

% of time spent translating	% of time spent summarizing
2%	0

Number of respondents	Number of respondents
1/28	0

TASKS INVOLVING LISTENING

Broadcasts

% of time spent translating	% of time spent summarizing
44.2%	73%

Number of respondents	Number of respondents
10/28	5/28

The most frequently-chosen category is "narcotics trafficking," Business/finance, economics, science/technology, military, and legal were chosen seldom, if at all. "Other" tasks include radio transmissions and ship-to-shore, ship-to-ship broadcasts.

Monitoring of Live Conversations

Telephone:

% of time spent translating	% of time spent summarizing
33.5%	25.6%

Number of respondents	Number of respondents
21/28	19/28

Categories most often chosen include theft/white collar crime, organized crime, narcotics trafficking, terrorism, and counterintelligence. The other categories were seldom chosen.

Body Microphone:

% of time spent translating	% of time spent summarizing
21.8%	30.6%

Number of respondents	Number of respondents
16/28	8/28

The item chosen most often is narcotics trafficking. The other items on the checklist were seldom chosen. "Other" responses included microphone surveillance of live monitoring, Title III Live monitoring, TIV, and room ("hidden") mikes.

Recorded Conversations

Telephone:

% of time spent translating	% of time spent summarizing
38.7%	60.9%

Number of respondents	Number of respondents
27/28	14/28

The items most frequently chosen are the same as those for live conversations. The individual participants seem to have a wider range of experience with recorded rather than live material.

Body Recorder:

% of time spent translating	% of time spent summarizing
25.0%	32.0%

Number of respondents	Number of respondents
26/28	9/28

Other: (Answers included pretext calls and consensual recordings)

% of time spent translating	% of time spent summarizing
9.0%	27.8%

Number of respondents	Number of respondents
6/28	4/28

SECOND QUESTIONNAIRE: QUESTIONNAIRE TO DETERMINE FBI'S TRANSLATION NEEDS

ORAL TASKS

Interpretation Assignments

Number of respondents:	18/28
% of time spent	5%

The category most often chosen is "unannounced visitors." A frequent category listed under "other" is listening to three-way phone calls. Other categories include field interviews of witnesses and polygraph examinations.

Oral Proficiency Test

Number of respondents:	1/28
% of time spent	4%

WRITTEN TASKS

Legal Documents

% of time spent translating	% of time spent summarizing
15%	10.5%

Number of respondents	Number of respondents
11/28	2/28

All categories were checked, but "extradition requests" was chosen very infrequently. "Other" categories listed include: police reports, depositions, foreign consulate reports, and statements.

Booklets/Manuals

% of time spent translating	% of time spent summarizing
11.3%	5%

Number of respondents	Number of respondents
6/28	1/28

"Training manuals" and "science/technology" were the items most often chosen.

Forms

% of time spent translating	% of time spent summarizing
18%	1%
Number of respondents	**Number of respondents**
3/28	2/28

"Bureau forms" was checked most often.

Other

% of time spent translating	% of time spent summarizing
3%	0
Number of respondents	**Number of respondents**
2/28	0

"Other" responses include correspondence and press releases.

APPENDIX R

RFP STATEMENT OF WORK

SECTION C - Description/Specs./Work Statement

A. The following requirements and goals must be met by the offeror:

1. Purpose:

   a. The developed translation test will be used to test the translations skills of individuals.

   b. Currently translation skills are tested by means of written tests, which are to be translated verbatim from the foreign language into English and from English into the foreign language. The various tests vary in difficulty as well as in form and type of content. Due to the test form and lack of clear, standardized scoring criteria, the scores tend to lack consistency and hence, reliability. the tests lack some content validity, because they fail to measure summary translation skills from audio stimuli.

   c. The contractor is to provide scoring criteria based on, and consistent with, the Interagency Language Roundtable (ILR) level descriptions, with a scale from 0 to 5. (See Attachment D for a copy of the ILR level descriptions for speaking, listening, reading, and writing.) The test should be constructed in such a way as to facilitate easy, but finely calibrated scoring, perhaps by means of specified point penalty for categories of errors, e.g. mistranslation, grammar, word choice, style, etc., with an exact easy to apply notation system, which would ultimately result in a score which can be converted to the 0 through 5 scale. A rating sheet to register error types and calibrations will be helpful for this purpose.

d. The developed translation test should consist of an audio stimulus to test summary translation skill up to level 3, to establish a floor, plus a written stimulus to test full, verbatim translation skills between levels 2+ and 5, to establish a ceiling. There should be at least one alternate version of the test for retesting purposes.

e. The contractor will be able to some extent draw on the expertise of the master translators in the FBI, and personnel from the FBI could also be used for the audio portions of the test if desired.

f. The desired output should include a model and alternate in English, and Spanish test plus an alternate, and possibly additional tests in other languages, all of which should have been field-tested to provide quantifiable data regarding reliability, validity, administrative ease and scorability.

g. Upon completion of the contract the contractor will provide written instructions for the grading of the tests and if necessary a training session.

h. All materials generated during the course of the research, including notes and rough drafts, are to be turned over to the FBI.

2. Deliverables

The following are required to be furnished:

a. Monthly progress reports

b. Translation skill level descriptions

c. Audio cassettes with oral recordings of stimuli and appropriate documentation:

    (1) one plus an alternate in English

    (2) one plus an alternate in Spanish

f. Hard copies of written stimuli and appropriate documentation:

    (1) one plus an alternate in English

    (2) one plus an alternate in Spanish

g. Grading procedures, rating sheets and appropriate training manual

h. Three days of training at FBI, 10th and Pennsylvania Avenue, N. W. Washington, D. C.

Appendix 16

END

U.S. Dept. of Education

Office of Education
Research and
Improvement (OERI)

ERIC

Date Filmed

March 29, 1991